

**UNIVERSIDADE FEDERAL DO ESPÍRITO SANTO
CENTRO DE CIÊNCIAS EXATAS
PROGRAMA DE PÓS-GRADUAÇÃO EM QUÍMICA**

**APRENDIZAGEM DE MÁQUINA NA SOLUÇÃO DE
PROBLEMAS QUÍMICOS: FLORESTA ALEATÓRIA
APLICADA À ESPECTROMETRIA NA REGIÃO DO
INFRAVERMELHO**

**MACHINE LEARNING TO SOLVE CHEMICAL PROBLEMS: RANDOM FOREST
APPLIED TO SPECTROMETRY IN THE INFRARED REGION**

MARCIA HELENA CASSAGO NASCIMENTO

Tese de Doutorado

**Vitória
2023**

MARCIA HELENA CASSAGO NASCIMENTO

**Aprendizagem de máquina na solução de problemas químicos:
floresta aleatória aplicada à espectrometria na região do
infravermelho**

Tese apresentada ao Programa de Pós-Graduação em Química do Centro de Ciências Exatas da Universidade Federal do Espírito Santo como requisito parcial para obtenção do Título de Doutora em Química.

Área de Concentração: Química

Linha de Pesquisa: Química do Petróleo e Biocombustíveis.

Orientador: Prof. Dr. Paulo Roberto Filgueiras

Coorientador: Prof. Dr. Wanderson Romão

VITÓRIA
2023

Ficha catalográfica disponibilizada pelo Sistema Integrado de
Bibliotecas - SIBI/UFES e elaborada pelo autor

N244a Nascimento, Marcia Helena Cassago, 1976-
Aprendizagem de máquina na solução de problemas químicos
: Floresta aleatória aplicada à espectrometria na região do
infravermelho / Marcia Helena Cassago Nascimento. - 2023.
214 f. : il.

Orientador: Paulo Roberto Filgueiras.
Coorientador: Wanderson Romão.
Tese (Doutorado em Química) - Universidade Federal do
Espírito Santo, Centro de Ciências Exatas.

1. Quimiometria. 2. Espectroscopia no infravermelho. 3.
Diesel. 4. Covid-19. 5. Saliva. 6. Aprendizado de máquina. I.
Filgueiras, Paulo Roberto. II. Romão, Wanderson. III.
Universidade Federal do Espírito Santo. Centro de Ciências
Exatas. IV. Título.

CDU: 54

Aprendizagem de máquina na solução de problemas químicos: floresta aleatória aplicada à espectrometria na região do infravermelho

Marcia Helena Cassago Nascimento

Tese submetida ao Programa de Pós-Graduação em Química do Centro de Ciências Exatas da Universidade Federal do Espírito Santo como requisito parcial para a obtenção do Grau de Doutor(a) em Química.

Documento assinado digitalmente
 PAULO ROBERTO FILGUEIRAS
Data: 14/02/2023 22:18-0300
Verifique em <https://verificador.iti.br>

Aprovada em 13/02/2023 por:

Documento assinado digitalmente
 WANDERSON ROMAO
Data: 15/02/2023 08:28:14-0300
Verifique em <https://verificador.iti.br>

Prof.(a) Dr.(a) Paulo Roberto Filgueiras¹
Universidade Federal do Espírito Santo
Orientador(a)

Documento assinado digitalmente
 MARCONE AUGUSTO LEAL DE OLIVEIRA
Data: 16/02/2023 08:12:45-0300
Verifique em <https://verificador.iti.br>

Prof.(a) Dr.(a) Wanderson Romão¹
Instituto Federal do Espírito Santo
Coorientador(a)

Mariana
Ramos de
Almeida


Assinado de forma
digital por Mariana
Ramos de Almeida
Dados: 2023.02.16
20:06:14 -05'00'

Documento assinado digitalmente
 MARCO FLORES FERRAO
Data: 17/02/2023 15:31:40-0300
Verifique em <https://verificador.iti.br>

Prof.(a) Dr.(a) Mariana Ramos Almeida¹
Universidade Federal de Minas Gerais

Prof.(a) Dr.(a) Marco Flores Ferrão¹
Universidade Federal do Rio Grande do Sul

Prof.(a) Dr.(a) Maria Tereza Weitzel Dias Carneiro Lima¹
Universidade Federal do Espírito Santo

¹ O documento será assinado eletronicamente em conformidade com as normas prescritas na Portaria Normativa PRPPG/ UFES nº 08/2021.

Universidade Federal do Espírito Santo
Vitória, fevereiro de 2023



UNIVERSIDADE FEDERAL DO ESPÍRITO SANTO

PROTOCOLO DE ASSINATURA



O documento acima foi assinado digitalmente com senha eletrônica através do Protocolo Web, conforme Portaria UFES nº 1.269 de 30/08/2018, por
MARIA TEREZA WEITZEL DIAS CARNEIRO LIMA - SIAPE 1116651
Departamento de Química - DQ/CCE
Em 17/02/2023 às 17:58

Para verificar as assinaturas e visualizar o documento original acesse o link:
<https://api.lepisma.ufes.br/arquivos-assinados/655313?tipoArquivo=O>

Ao meu filho Luiz Felype, minha maior inspiração, ao meu neto Zyon, minha alegria, e à
minha mãe, dedico este trabalho em nome de todo sacrifício e apoio.

AGRADECIMENTOS

Existem muitas pessoas as quais gostaria de agradecer desde o dia que ingressei à Universidade Federal do Espírito Santo, há aproximadamente sete anos, com muita insegurança e timidez. Tudo aqui era para mim um sonho distante, ao qual muitas vezes fui levada a acreditar que nunca poderia ser alcançado.

Gostaria de agradecer primeiramente a Deus! Eu acredito que Ele tem me fortalecido e sustentado em todos os aspectos. Além disso, costumo dizer que Deus coloca anjos em torno de mim a cada etapa da minha vida, aqui não foi diferente.

Gostaria de agradecer à minha família, principalmente minha mãe pela total cumplicidade, me apoiando e sendo meu porto seguro, mesmo sem entender o porquê de tanto sacrifício, e às vezes até soltando um “eu preferia que você só trabalhasse na escola”. Estendo este meu agradecimento ao meu filho, que sempre me inspira, minhas sobrinhas e sobrinhos que entendem minha ausência e às vezes ficam lá do meu lado, só me fazendo companhia, enquanto estava perdida em minha concentração.

Agradeço ao Márcio, meu namorado, por toda compreensão e apoio ao longo desses sete anos de minha jornada desde o mestrado.

Gostaria de fazer um agradecimento que antecede minha jornada na UFES, aos professores que tive na educação básica por quem fui abraçada de uma forma que não sei explicar, muitos dos quais me fizeram sonhar e sem essa chama de um sonho eu jamais estaria hoje aqui escrevendo esta tese. Aos professores do Curso Técnico em Química, em especial ao Prof. Me. Arnaldo Costa Barros Filho, que propulsionou minha história de superação e me acompanhou até a graduação. Aos professores da Graduação em Química na Faculdade FAESA. E aqui aproveito para agradecer enormemente às políticas públicas para educação, pelas quais pude ingressar no Curso Técnico em Química e chegar até este Doutorado em Química. Sim, sem essas políticas que ajudam a democratizar o acesso ao ensino, eu não conseguiria ter percorrido esse caminho, tenho total consciência disso. Por isso insisto: Viva às políticas públicas de acesso à educação!

Gostaria de agradecer as pessoas que me ensinaram muito, em um ambiente de trabalho que se tornou para mim referência de conhecimento e respeito, Dr. Eduardo Batitucci, sua esposa Dr.^a Maria Hermenegilda e sua família. Com certeza muito de minha formação foi dada no tempo em que pude conviver e aprender com vocês. Aí também se iniciou minha aproximação com a UFES, até então muito distante para mim. Flávio e Eliane (*in memoriam*), com gestos singelos como dar de presente uma calculadora e emprestar um livro de cálculo, ao me verem ingressar na faculdade. Gestos que trazem um significado importante nos momentos de desafios.

Como disse anteriormente, Deus vive colocando anjos em minha vida e aqui no laboratório de Quimiometria da UFES convivi com vários, que me atendem sempre com muito carinho. Acredito que um deles veio na pessoa da Betina, com toda sua firmeza quando eu precisava, apoio, ensinamentos e amizade. Obrigada! Aos que ainda estão aqui caminhando comigo, Gabriely, Pedro, Bárbara, Bruno e Madson. E aos que já seguiram outros rumos, Ellisson, Francine, Layla, Mariana, Andressa, Thales, Matheus, Érica e Rayza. Obrigada!

Gostaria de agradecer a todos os professores do Programa de Pós-Graduação em Química da UFES, todos os conhecimentos agregados, aos que fiz disciplinas diretamente e até aos que pude somente acompanhar, conhecer, tudo isso é um rico aprendizado para mim. Em especial à Prof.^a Dr^a Geisamanda Pedrini e ao Prof. Dr. Rafael Queiroz, com os quais pude fazer os estágios docentes, pelo rico aprendizado e convívio.

Gostaria de agradecer ao Prof. Dr. Valério Barauna e toda sua equipe do Programa de Pós-Graduação em Fisiologia da UFES, pelos trabalhos colaborativos que tem sido para mim um desafio importante, mas mais do que isso, tem me oferecido a oportunidade de conhecer novos amigos.

Gostaria de deixar aqui registrado meu singelo agradecimento para Prof. Dr. Steven D. Brown (*Dept. of Chemistry and Biochemistry, University of Delaware, Newark, US*) e Prof. Dr. Roy Edward Bruns – UNICAMP, que gentilmente autorizaram a utilização de materiais extraídos de suas palestras para que eu pudesse compor minha tese.

Agradeço ao meu orientador, Prof. Dr. Paulo Roberto Filgueiras, pelos ensinamentos e por toda orientação, por ter acreditado em mim desde o dia em que o abordei com pedido de sua orientação para o Mestrado, também lhe agradeço pela sua paciência e compreensão.

Agradeço ao meu coorientador, Prof. Dr. Wanderson Romão, por todo conhecimento, orientação e apoio. Também o agradeço pela oportunidade de conviver em um grupo de pesquisa com tantas realizações e importância. Aproveito para agradecer a todos os amigos que tenho feito também neste grupo.

Gostaria de agradecer ao Núcleo de Competências em Química do Petróleo, nas pessoas do Prof. Dr. Eustáquio e todos os funcionários, à Suzy, a Dona Maria, ao Oscar e todos os demais. Estendo este agradecimento a todas as equipes dos conjuntos de laboratórios do Núcleo, ao coordenador do PPGQUI, Prof. Dr. Cleocir Dalmaschio e aos funcionários, incluindo Alex, que mesmo hoje em outro setor, deixou sua marca e seu carisma em nossas lembranças e corações.

Meu agradecimento aos membros da banca de Qualificação, Profa. Dra. Maria Tereza Weitzel Dias Carneiro Lima, Profa. Dra. Mariana Ramos Almeida e Profa. Dra. Betina Pires Oliveira, por todas as ponderações e contribuições que estão inseridas nesta tese. Assim como agradeço aos professores membros da banca para esta defesa, Profa. Dra. Maria Tereza Weitzel Dias Carneiro Lima; Profa. Dra. Mariana Ramos Almeida; Prof. Dr. Marccone Augusto Leal de Oliveira e Prof. Dr. Marco Flores Ferrão. Obrigada!

Por fim, gostaria de agradecer às agências de fomento à pesquisa, CNPq, CAPES e FAPES pelo incentivo a pesquisa e apoio financeiro na forma de bolsa de doutorado.

Novamente elevo meus agradecimentos a Deus, por me permitir concluir esta etapa de minha vida acadêmica e pela alegria renovada com a chegada de meu netinho Zyon, nascido em 2022.

A vida não é fácil para nenhum de nós. Mas e daí? Precisamos ter perseverança e, acima de tudo, autoconfiança. Precisamos acreditar que temos talento para alguma coisa e que essa coisa deve ser alcançada.

Marie Curie.

LISTA DE FIGURAS

Figura 2.1 Prof. Dr. Roy Edward Bruns demonstrando fita magnética, em web conferência para mesa redonda em XI Workshop de Quimiometria. ⁴⁶	29
Figura 2.2. Cartão perfurado para análise de dados de pesquisadores do Instituto de Química (A) e computador PDP10 da marca Digital, pertencente à Universidade de Campinas - UNICAMP (B), onde eram realizadas as análises dos dados.....	30
Figura 2.3 Capa de edição dedicada ao Prof. Dr. Roney Jesus Poppi, periódico científico Brazilian Journal of Analytical Chemistry, em 2021. Fonte: BrJAC, volume 8, n.32 (2021) ⁵⁵ . 31	
Figura 2.4 Esquema geral de etapas iterativas para construção e aplicação de modelos de aprendizagem de máquina a dados químicos.....	34
Figura 2.5 Número total de artigos publicados, por ano, a partir de análise bibliométrica nas bases <i>Web of Science</i> e <i>Scopus</i> sobre a associação de RF e espectroscopia nos anos de 2018 a 2023 (15/01/2023).....	40
Figura 2.6 Dez fontes mais relevantes em número de documentos (A) e total de citações (B), a partir de análise bibliométrica nas bases <i>Web of Science</i> e <i>Scopus</i> sobre a associação de RF e espectroscopia nos anos de 2018 a 2023 (15/01/2023).	41
Figura 2.7 Produção e contribuições entre países com associação dos termos floresta aleatória e espectroscopia no período de 2018 (01/01/2018) a 2023 (15/01/2023). (A) Dez países mais produtivos (SCP: produções sem contribuições entre países; MCP: produções com contribuição entre países); contribuição e com contribuição; (B) Mapa de colaboração entre os países.	42
Figura 2.8 Um esquema geral de uma árvore de decisão.	49
Figura 2.9 Exemplo esquemático de divisão de uma árvore de decisão.	50
Figura 2.10 Estrutura básica de métodos ensemble.....	51
Figura 2.11 Esquema simplificado de etapas de MPA para classificação. Fonte: adaptado de Li et al. ¹⁶⁹	53
Figura 2.12 Fluxograma representando estruturas possíveis de MPA. Fonte: adaptado de Deng et al. ¹⁶⁸	54
Figura 2.13 Fluxograma de crescimento de árvore de classificação ou de regressão em modelo de RF.....	56
Figura 2.14 Fluxograma com passos para construção de modelo de floresta aleatório de classe única (OCRF). Fonte: compilado pela autora.	59
Figura 2.15 Fluxograma para construção de modelo floresta aleatória não-supervisionado (URF).	62
Figura 3.1 Avaliação univariada dos métodos de ensaio das propriedades físico-químicas do óleo diesel metropolitano S500 (Δ, DM) e óleo diesel marítimo A (O,DMA). (a) ponto de fulgor; (b) ponto de fluidez; (c) densidade; (d) viscosidade cinemática; (e) NAT; (f) teor de enxofre.....	95
Figura 3.2 Espectros normalizados (A) de 48 amostras de diesel de estudo de caso. Bandas características do perfil espectral de diesel: 3000 - 2800 cm ⁻¹ (B); 1500 - 1300 cm ⁻¹ (C); e 900 - 650 cm ⁻¹ (D).....	97
Figura 3.3 Espectro médio das amostras de diesel do caso de estudo identificadas como amostras conformes e amostras não conforme para o parâmetro TAN (ANP). (A) espectro completo; (B) região destacada da impressão digital de 1000 a 650 cm ⁻¹ ; (C) região destacada de 1800 a 1600 cm ⁻¹	98

Figura 3.4 Espectros MIR na região de 650 - 4000 cm^{-1} de (A) amostra de diesel puro adotada como referência (DP); (B) terebintina (T); (C) óleo de soja (S).....	99
Figura 3.5 Gráficos de <i>scores</i> (A) e <i>loadings</i> (B – D) de modelo PCA (PC1 versus PC2 e PC3) de espectros MIR de 48 amostras de diesel do estudo de caso, a partir das variáveis selecionadas de 1816 a 1650 cm^{-1}	100
Figura 3.6 Gráfico PCoA das coordenadas (distância Euclidiana) a partir da matriz de dissimilaridade de 48 amostras de diesel do caso de estudo no modelo URF além das misturas com terebintina (Adulteração 1)) e óleo de soja (Adulteração 2).	102
Figura 4.1 Diagrama descritivo da metodologia deste estudo, contendo as etapas de admissão dos pacientes e coleta do biofluidos, aquisição dos dados e análise multivariada dos dados.	116
Figura 4.2 Espectros MIR médios normalizados do conjunto de dados de saliva (n=237) de pacientes do coorte deste estudo (linha azul: espectro médio do conjunto de amostras negativas para COVID-19, n=99); linha vermelha: espectro médio do conjunto de amostras positivas COVID-19, n=138).	125
Figura 4.3 Análise de coordenadas principais (PCoA) a partir de modelo URF de conjunto de dados espectros MIR de saliva (n=246).	127
Figura 4.4. Variáveis mais importantes (coeficiente PLS-DA) e selecionadas por meio de métodos de seleção de variáveis e modelos de classificação. (A) Variáveis selecionadas pelo método GA-LDA; (B) variáveis selecionadas pelo método SPA-LDA; (C) Variáveis selecionadas pelo método PSO; (D) Variáveis mais importante para classe 1 (+ vermelha) e classe 2 (+ azul) por meio de análise dos valores de coeficientes PLS_DA; (E) Variáveis mais importante para classe 1 (+ vermelha) e classe 2 (+ azul) por meio da análise dos valores de coeficientes PSO-PLS-DA.	130
Figura 4.5 Gráfico das classes real e prevista e dias dos primeiros sintomas do conjunto de dados (n=237). (A) Marcadores de classe positivos; (B) Marcadores de classe negativos.	132
Figura 4.6. Distribuição de parâmetro F1 Score de classe 1 (Positivo) de todos os modelos construídos no teste de permutação de y (histogramas), em comparação com F1 Score de modelos construídos neste estudo (linha vermelha na vertical), para conjunto de treinamento e de teste.	133
Figura 5.1 Dois tipos de swab de rayon observados neste estudo, sendo utilizados para coleta nasofaríngea para diagnóstico de COVID-19. (A) Swab normal; (B) swab fino.....	146
Figura 5.2 Posicionamento no equipamento, dos <i>swab</i> utilizados neste estudo, para evitar contato do material da haste com o cristal ATR. (A) Swab normal; (B) Swab fino.....	146
Figura 5.3 Fluxograma de aplicação de URF para selecionar variáveis preditoras.	149
Figura 5.4 Fluxograma de aplicação de URF para classificação de amostras.....	150
Figura 5.5 Espectros adquiridos em diferentes posições dos swab utilizados neste estudo, sem coleta de material biológico. A e B: Swab de rayon, fino; C e D: Swab de rayon, normal (maior espessura e maior quantidade de fibras; E e F: Swab de algodão. Espectros brutos (A, C e E) <i>versus</i> espectros após correção de linha de base (B, D e F).....	152
Figura 5.6 Gráficos de análises exploratórias com dados originais sem subtração dos espectros de swab. A) Gráfico de Scores PC1 x PC2 de modelo PCA (dados com quadruplicatas, processados com airPLS e centrado na média); B) Gráfico PCoA de URF-PCoA PCo1 x PCo2 (dados com espectros médios, processados com airPLS, MSC e primeira derivada, 7 pontos).	153
Figura 5.7 Espectros dos conjuntos de dados após pré-processamentos iniciais (ajuste de linha de base, subtração de espectro, suavização SG). A) Conjunto de amostras <u>Positivo</u> COVID-19, <u>Swab fino com secagem</u> (n=71); B) Conjunto de amostras <u>Negativo</u> COVID-19,	

Swab fino com secagem (n=80); C) Conjunto de amostras Positivo COVID-19, Swab fino sem secagem (n=90); D) Conjunto de amostras Negativo COVID-19, Swab fino sem secagem (n=100); E) Conjunto de amostras Positivo COVID-19, Swab normal com secagem (n=61); F) Conjunto de amostras Negativo COVID-19, Swab normal com secagem (n=11); G) Conjunto de amostras Positivo COVID-19, Swab normal sem secagem (n=66); H) Conjunto de amostras Negativo COVID-19, Swab normal sem secagem (n=60)..... 154

Figura 5.8 Espectros médios Positivo (em vermelho) e Negativo (em cinza) dos conjuntos de dados após pré-processamentos iniciais (ajuste de linha de base, subtração de espectro, suavização SG). A) Conjunto swab fino com secagem; B) Conjunto swab fino sem secagem; C) Conjunto swab normal com secagem; D) Conjunto swab normal sem secagem. 155

Figura 5.9 Gráficos de coordenadas principais, a partir de modelo URF-PCoA (distância Euclidiana) de conjuntos de dados avaliados. A) Conjunto swab fino com secagem; B) Conjunto swab fino sem secagem; C) Conjunto swab normal com secagem; D) Conjunto swab normal sem secagem. 156

Figura 5.10 Gráficos de coordenadas principais, a partir de modelo URF-PCoA (distância Euclidiana) de conjuntos de dados avaliados. A) Conjunto swab fino + swab normal com secagem; B) Conjunto swab fino + swab normal sem secagem. 157

Figura 5.11 Demonstração gráfica da frequência relativa para identificação de variáveis preditoras, a partir de dois métodos de identificação: A) modelo URF-PCoA amostras não rotuladas (conjunto swab fino com secagem); B) modelo URF-PCoA amostras rotuladas (conjunto swab fino com secagem). 159

Figura 5.12 Demonstração gráfica da frequência relativa para identificação de variáveis preditoras, a partir de dois métodos de identificação: A) modelo URF-PCoA amostras não rotuladas (conjunto swab fino sem secagem); B) modelo URF-PCoA amostras rotuladas (conjunto swab fino sem secagem). 159

LISTA DE TABELAS

Tabela 2.1 Ocorrência de termos-chave a partir de uma pesquisa qualitativa nos resultados de bibliometria com termos-chave “Random forest” AND “Spectroscop*”	38
Tabela 2.2 Amostragem de estudos associando técnicas espectroscópicas a modelos RF.	39
Tabela 2.3. Demonstração de Tabela de contingência em problema de duas classes.	63
Tabela 2.4 Demonstração de Tabela de contingência em problema hipotético de classificação de 90 amostras de biofluido positivo para determinada doença e 100 amostras negativas.	65
Tabela 2.5 Principais parâmetros de desempenho para validação de modelos de classificação.	67
Tabela 3.1 Características de desempenho para avaliação de modelos de classificação ...	92
Tabela 4.1 Configurações dos modelos.....	122
Tabela 4.2 Dados gênero e comorbidades do coorte do estudo.	124
Tabela 4.3 Principais atribuições de bandas presentes nos espectros. ^{6,8}	126
Tabela 4.4 Matriz de confusão de modelos de classificação de consenso.	128
Tabela 4.5 Parâmetros de avaliação de desempenho de modelos de classificação de conjuntos de treinamento (linha sem sombreado) e teste (linhas sombreadas).	129
Tabela 5.1 Parâmetros de desempenho de modelos de classificação URF-PCoA-RF, para a classe Positivo, de conjuntos de treinamento e teste (linha sombreada).....	160

LISTA DE ABREVIATURAS E SIGLAS

ANN	Redes neurais artificiais (do inglês, <i>Artificial neural networks</i>)
ANP	Agência Nacional do Petróleo, Gás Natural e Biocombustíveis
API	Instituto americano de petróleo (do inglês, <i>American Petroleum Institute</i>)
ASTM	Sociedade americana de testes e materiais (do inglês, <i>American Society for Testing and Materials</i>)
ATR	Reflectância total atenuada (do inglês, <i>Attenuated total reflectance</i>)
Bagging	Agregação pelo <i>bootstrap</i> (do inglês, <i>Bootstrapping aggregation</i>)
CAPES	Coordenação de Aperfeiçoamento de Pessoal de Nível Superior
CART	Árvore de classificação e de regressão (do inglês, <i>Classification and regression tree</i>)
CNPq	Conselho Nacional de Desenvolvimento Científico e Tecnológico
COVID-19	Doença por coronavírus 2019 (do inglês, <i>Coronavirus disease 2019</i>)
CT	Número de ciclos (do inglês, <i>Cycle threshold</i>)
CTC	Capacidade de troca de cátions
DNA	Ácido desoxirribonucleico (do inglês, <i>Deoxyribonucleic acid</i>)
DPOC	Doença pulmonar obstrutiva crônica
ELISA	Ensaio de imunoabsorção enzimática (do inglês, <i>Enzyme-Linked Immunosorbent Assay</i>)
ENSEMBLE	Sistema de classificadores múltiplos (do inglês, <i>Multiple classifier systems</i> , ou, <i>committee-based learning</i>)
ER	Taxa de erro (do inglês, <i>Error rate</i>)
ES	Espírito Santo
ESPEC	Especificidade
EXAT	Exatidão
FAPES	Fundação de Amparo À Pesquisa do Estado do Espírito Santo
FN	Número de Falsos Negativos
FP	Número de Falsos Positivos
FPR	Taxa de Falso positivo (do inglês, <i>False positive rate</i>)
FTIR	Infravermelho com transformada de Fourier (do inglês, <i>Fourier transform infrared</i>)
GA	Algoritmo genético (do inglês, <i>Genetic algorithm</i>)
GA-LDA	Análise discriminante pelo algoritmo genético (do inglês, <i>Genetic algorithm linear with discriminant analysis</i>)
HATR	espectroscopia com reflectância total atenuada horizontal (do inglês, <i>Horizontal attenuated total reflectance</i>)
HCA	Análise hierárquica de cluster (do inglês, <i>Hierarchical cluster analysis</i>)
iCEPi	Instituto Capixaba de Ensino, Pesquisa e Inovação em Saúde
In-Bag	Amostras de treinamento “na sacola” (do inglês, <i>In-Bag samples</i>)

iPCA	Análise de componentes principais por intervalo (do inglês, <i>Interval principal component analysis</i>)
IR	Infravermelho (do inglês, <i>Infrared</i>)
IUPAC	União Internacional de Química pura e aplicada (do inglês, <i>International union of pure and applied Chemistry</i>)
KOH	Hidróxido de potássio
LABPETRO	Laboratório de Pesquisa e Desenvolvimento de Metodologias para Análise de Petróleos
LACEN	Laboratório Central
LDA	Análise discriminante linear (do inglês, <i>Linear discriminant analysis</i>)
LIBS	Espectrometria de emissão óptica com plasma induzido por laser (do inglês, <i>Laser induced-breakdown spectroscopy</i>)
MCC	Coefficiente de correlação de Matthews (do inglês, <i>Matthews correlation coefficient</i>)
MCR-ALS	Resolução multivariada de curvas com mínimos quadrados alternantes (do inglês, <i>Multivariate curve resolution with alternating least squares</i>)
MD	Diesel metropolitano (do inglês, <i>Metropolitan diesel</i>)
MDA	Diesel marítimo tipo A (do inglês, <i>Type A marine diesel</i>)
MDS	Dimensionamento multidimensional (do inglês, <i>Multidimensional scaling</i>)
MERS	Síndrome respiratória do Oriente Médio (do inglês, <i>Middle east respiratory syndrome</i>)
MIR	Infravermelho na região do médio (do inglês, <i>Mid-infrared region</i>)
MOS	Matéria orgânica do solo
MPA	Análise de população de modelos (do inglês, <i>Model population analysis</i>),
MSC	correção de dispersão multiplicativa (do inglês, <i>Multiplicative scatter correction</i>)
NCQP	Núcleo de Competências em Química do Petróleo
NEG	Negativo
NIR	Espectroscopia na região do infravermelho próximo (do inglês, <i>Near infrared region</i>)
OCRF	Floresta aleatória de classe única (do inglês, <i>One-class random forest</i>)
OMS	Organização Mundial de Saúde
OOB	Amostras de teste “fora da sacola” (do inglês, <i>Out-of-Bag</i>)
PAH	Hidrocarboneto aromático policíclico (do inglês, <i>Polycyclic aromatic hydrocarbon</i>)
PARAFAC	Análise de fatores paralelos (do inglês, <i>Parallel Factor analysis</i>)
PC	Componente principal (do inglês, <i>Principal component</i>)
PCA	Análise de componentes principais (do inglês, <i>Principal component analysis</i>)
PCo	Coordenada principal (do inglês, <i>Principal coordinate</i>)
PCoA	Análise de coordenadas principais (do inglês, <i>Principal cordinate analysis</i>)
PLS	mínimos quadrados parciais (do inglês, <i>Partial least squares</i>)

PLS-DA	Análise discriminante pelos mínimos quadrados parciais (do inglês, <i>Partial least squares-discriminant analysis</i>)
POS	Positivo
PREC	Precisão
PSO	Otimização de enxame de partículas (do inglês, <i>particle swarm optimization</i>)
PTR-MS	Espectrometria de massas de reação de transferência de prótons (do inglês, <i>Proton-transfer-reaction mass spectrometry</i>)
PXRF	Fluorescência de raio X portátil (do inglês, <i>Portable X-ray fluorescence spectrometer</i>)
RAM	Memória de acesso aleatório (do inglês, <i>Random access memory</i>)
RF	Floresta aleatória (do inglês, <i>Random forest</i>)
RF-RFE	árvore aleatória com eliminação recursiva de variáveis (do inglês, <i>Random forest-recursive feature elimination</i>)
RMN	Ressonância magnética nuclear
RNA	Ácido ribonucleico (do inglês, <i>RiboNucleic Acid</i>)
RT-qPCR	Reação em cadeia de polimerase de transcrição reversa (do inglês, <i>Quantitative reverse transcription Polymerase Chain Reaction</i>)
SARS	Síndrome respiratória severa aguda (do inglês, <i>Severe acute respiratory syndrome</i>)
SARS-CoV-2	Coronavírus da síndrome respiratória aguda grave 2 (do inglês, <i>Severe acute respiratory syndrome coronavirus 2</i>)
SB	Soma das bases
SENS	Sensibilidade
SERS-Raman	Espectroscopia Raman intensificado por superfície (do inglês, <i>Surface-enhanced Raman scattering</i>)
SESA	Secretaria de Estado da Saúde do Espírito Santo
SG	Savitzky-Golay
SIMCA	Modelagem flexível independente por analogia de classe (do inglês, <i>Soft Independent Modeling of Class Analogy</i>)
SNV	padronização da variação de sinais (do inglês, <i>standard normal variate</i>)
SPA-LDA	Análise discriminante pelo algoritmo de projeção sucessiva (do inglês, <i>Successive projection algorithm-linear discriminant analysis</i>)
SVD	decomposição por valores singulares (do inglês, <i>singular value decomposition</i>)
SVM	Máquina de vetores de suporte (do inglês, <i>Support vector machine</i>)
TAN	Número total de acidez (do inglês, <i>Total number of acidity</i>)
TFN	Taxa de Falso Negativo
TFP	Taxa de Falso Positivo
TNR	Taxa de Verdadeiro Negativo (do inglês, <i>True negative rate</i>)
TPR	Taxa de Verdadeiro Positivo (do inglês, <i>True positive rate</i>)
UFES	Universidade Federal do Espírito Santo

UNICAMP	Universidade Estadual de Campinas
URF	Floresta aleatória não supervisionada (do inglês, <i>Unsupervised random forest</i>)
UV-VIS	Espectroscopia na região ultravioleta e visível (do inglês, <i>Ultraviolet-visible spectrometry</i>)
VIS-NIR	Espectroscopia de reflectância na região do infravermelho próximo e do visível (do inglês, <i>Ultraviolet-visible-near infrared spectrophotometry</i>)
VL	Variáveis latentes
VN	Número de Verdadeiros Negativos
VP	Número de Verdadeiros Positivos
XAS	Espectroscopia de absorção de raios X (do inglês, <i>X-ray absorption spectroscopy</i>)

LISTA DE SÍMBOLOS

–	Ligação covalente simples
=	Ligação covalente dupla
%	Porcentagem
% v/v	Porcentagem em volume
wt%	Porcentagem em massa
H_0	Hipótese nula a ser testada em teste estatístico de hipóteses
H_a	Hipótese alternativa em teste estatístico de hipóteses
χ^2	Qui-quadrado
°C	Graus Celsius
C	Carbono
H	Hidrogênio
N	Nitrogênio
O	Oxigênio
P	Fósforo
S	Enxofre
cm^{-1}	Centímetro recíproco
g cm^{-3}	Grama por centímetro cúbico
kPa	Quilopascal
MB	Megabyte
mL	mililitros
n	Número de amostras
nm	Nanometro
R^2	Coeficiente de determinação
ν_{as}	estiramento assimétrico
ν_s	estiramento simétrico

δ deformação angular no plano

$mtry$ Parâmetro de variáveis selecionadas aleatoriamente

$nTrees$ Parâmetro de número total de árvores para modelo RF

α Alfa

β Beta

RESUMO

A Quimiometria iniciou na década de 70 com divulgação de estudos intitulados “*Computerized Learning Machines Applied to Chemical Problems*”, que expressam a motivação do surgimento deste campo de estudo: a necessidade de métodos multivariados desenvolvidos por químicos para resolver problemas químicos. Ao longo de 54 anos essa área tem se expandido e apresentado soluções para dados cada vez mais complexos, gerados pela Química Analítica moderna. Dentre os métodos de aprendizagem de máquina adaptados a problemas do ponto de vista químico, este trabalho buscou contribuir para maior compreensão, adaptação e aplicação do método floresta aleatória (RF). Este método faz parte de um grupo de métodos do tipo *ensemble*, que são caracterizados pelo agrupamento de classificadores gerando uma resposta combinada. RF pode ser utilizado como modelo de calibração multivariada, ou de reconhecimento de padrões, sendo este último o foco desta tese. Além do contexto histórico, descrevemos neste trabalho as adaptações que têm sido propostas para aplicação deste método nas soluções de problemas em Química, com diferentes técnicas analíticas e abordagens. Neste trabalho, RF foi aplicado para reconhecimento de padrões não supervisionado em estudo de caso de suspeita de fraude de combustíveis, como um método de triagem de amostras de diesel submetidas à espectroscopia na região do infravermelho médio (MIR). A interpretação do resultado da URF por meio de um gráfico de coordenadas principais (PCoA) permitiu a triagem de amostras cuja adulteração foi comprovada pelos ensaios de parâmetros físico-químicos. Ademais, adaptamos e aplicamos o método URF para contribuir em um outro campo de estudo, a biospectroscopia. Uma grande parte dos estudos neste campo buscam desenvolver métodos de diagnóstico, ou biópsia líquida, por meio de biofluidos e espectroscopia de vibração associada aos métodos quimiométricos para extrair a informação das mudanças bioquímicas provocadas pela doença ou pelo agente infeccioso. Adaptamos e aplicamos o método URF para identificar e modelar estruturas discriminantes em dados espectroscópicos em dois estudos neste campo: diagnóstico não invasivo de COVID-19 a partir de amostras de saliva analisadas pela espectroscopia MIR e uma proposta para reconhecimento de padrões e diagnóstico de COVID-19 a partir de swab nasofaríngeo e espectroscopia MIR. No primeiro, uma combinação de modelos de classificação distinguiu amostras de saliva de pessoas infectadas pela COVID-19 com exatidão de 85%, sensibilidade de 93% e especificidade de 74%. No último, URF foi utilizada de forma mais abrangente e inovadora, sendo ponto de partida para seleção de variáveis relevantes associadas a alterações de biomoléculas presentes nos biofluidos e serviu como dados de entrada para modelos de classificação. Com o uso da URF como dados entrada para modelos de classificação, as amostras de biofluidos coletados com dois tipos de swab foram classificadas com 87,6% de exatidão, 93,6% de sensibilidade, 79,4% de especificidade e F-Score de 0,898. As diferentes abordagens deste trabalho contribuem para divulgar a versatilidade e a eficiência do método RF, além de inovar sua adaptação, aproveitando a potencialidade deste método para os campos de estudo abordados.

Palavras-chave: Quimiometria; MIR; Diesel; URF; COVID-19; Saliva; Seleção de variáveis;

ABSTRACT

Chemometrics began in the 1970s with the publication of a series of studies entitled "Computerized Learning Applied Machines to Chemical Problems" which express the motivation for the emergence of this field of study: the need for multivariate methods developed by chemists to solve chemical problems. Over 54 years, this area has expanded and presented solutions for increasingly complex data generated by modern Analytical Chemistry. Among the machine learning methods adapted to problems from the chemical point of view, this study contributes to a greater understanding, adaptation, and application of the random forest (RF) method. It is an ensemble-based method of learning multiple classifier systems. RF can be as a multivariate calibration model or pattern recognition, the latter being the focus of this thesis. In addition to the historical context, we describe adaptations proposed for the RF method to solve Chemistry problems with different analytical techniques and approaches. In this study, we applied RF for unsupervised pattern recognition as a screening method in a case study of suspected fuel fraud of diesel samples submitted to Fourier transform spectroscopy in the mid-infrared region (FT-MIR). The interpretation of the URF through a principal coordinate graph (PCoA) allowed the screening of samples with adulteration confirmed by the test of physical-chemical parameters. In addition, we adapted and applied the URF method to contribute to another field of study: biospectroscopy. A large part of the studies in this field is to develop alternative diagnosis methods or liquid biopsy. It is possible through biofluids, and spectroscopy associated with chemometric methods to extract information from biochemical changes caused by the disease or infectious agent. We adapted URF to identify a discriminant structure in spectroscopic data from two studies: a noninvasive diagnosis of COVID-19 from saliva samples analyzed by FT-MIR, and a proposal for pattern recognition and diagnosis of COVID-19 from nasopharyngeal swab and FT-MIR. In the first, an ensemble of classification models distinguished saliva samples from COVID-19-infected people with an accuracy of 85%, a sensitivity of 93%, and a specificity of 74%. In another, URF was a comprehensive and innovative way: a starting point for selecting relevant variables and input data for classification models. With the URF as input data for classification models, we classified biofluid samples collected with two types of swabs with 87.6% accuracy, 93.6% sensitivity, 79.4% specificity, and 0.898 F-Score. Different approaches in this study contribute to disseminating the versatility and efficiency of the RF method, in addition to innovating its adaptation, taking advantage of the potential of this method for the different problems addressed.

Keywords: Chemometrics; MIR; Diesel fuel; URF; COVID-19; Saliva; variable selection.

SUMÁRIO

1. INTRODUÇÃO	20
1.1. OBJETIVOS	23
1.2. ESTRUTURA DA TESE	24
2. FLORESTA ALEATÓRIA APLICADA À ESPECTROMETRIA NA REGIÃO DO INFRAVERMELHO	20
2.1. PRINCÍPIO HISTÓRICO DA QUIMIOMETRIA.....	26
2.2. RECONHECIMENTO DE PADRÕES E APRENDIZAGEM DE MÁQUINAS	32
2.3. FLORESTA ALEATÓRIA APLICADA À ESPECTROSCOPIA QUÍMICA	35
2.4. FUNDAMENTAÇÃO TEÓRICA.....	48
2.5. REFERÊNCIAS.....	67
3. O USO DE FLORESTA ALEATÓRIA NÃO SUPERVISIONADA NA DISCRIMINAÇÃO DO ÓLEO DIESEL: UM ESTUDO DE CASO DE ADULTERAÇÃO DE COMBUSTÍVEL	84
3.1. INTRODUÇÃO	85
3.2. METODOLOGIA.....	87
3.3. RESULTADOS E DISCUSSÃO	92
3.4. CONCLUSÕES	104
3.5. REFERÊNCIAS.....	105
4. DIAGNÓSTICO NÃO INVASIVO PARA COVID-19 VIA ESPECTROSCOPIA FTIR E ANÁLISE MULTIVARIADA A PARTIR DE SALIVA.....	112
4.1. INTRODUÇÃO	113
4.2. METODOLOGIA.....	115
4.3. RESULTADOS E DISCUSSÃO	123
4.4. CONCLUSÕES	134
4.5. REFERÊNCIAS.....	135
5. FLORESTA ALEATÓRIA NÃO SUPERVISIONADA COMO ALTERNATIVA PARA RECONHECIMENTO DE PADRÕES E DIAGNÓSTICO DE COVID-19 A PARTIR DE SWAB NASOFARÍNGEO E ESPECTROSCOPIA MIR	141
5.1. INTRODUÇÃO	142
5.2. METODOLOGIA.....	144
5.3. RESULTADOS E DISCUSSÃO	151
5.4. CONCLUSÕES	163
5.5. REFERÊNCIAS.....	164
6. CONCLUSÕES DA TESE	169
7. MATERIAL SUPLEMENTAR E ANEXOS.....	171
7.1. MATERIAL SUPLEMENTAR - O USO DE FLORESTA ALEATÓRIA NÃO SUPERVISIONADA NA DISCRIMINAÇÃO DO ÓLEO DIESEL: UM ESTUDO DE CASO DE ADULTERAÇÃO DE COMBUSTÍVEL (CAPÍTULO 3).....	171
7.2. MATERIAL SUPLEMENTAR – DIAGNÓSTICO NÃO INVASIVO PARA COVID-19 A PARTIR DE BIOFLUIDO SALIVA VIA ESPECTROSCOPIA FTIR E ANÁLISE MULTIVARIADA (CAP. 04)	177
7.3. MATERIAL SUPLEMENTAR - FLORESTA ALEATÓRIA NÃO SUPERVISIONADA COMO ALTERNATIVA PARA RECONHECIMENTO DE PADRÕES E DIAGNÓSTICO DE COVID-19 A PARTIR DE SWAB NASOFARÍNGEO E ESPECTROSCOPIA MIR (CAP. 05).....	190
7.4. ANEXO A - EXEMPLO DE CÁLCULO DE ÍNDICE DE GINI EM ÁRVORE DE CLASSIFICAÇÃO.....	199
7.5. PRODUÇÃO CIENTÍFICA	202

1. INTRODUÇÃO

Uma importante relação entre Química e ciências de dados tem sido presente desde a década de 70, quando a computação científica se tornou de certa forma acessível, permitindo conectividade de equipamentos a computadores.¹ Esse avanço nos métodos de análise propiciou um aumento considerável na geração de dados analíticos de diferentes complexidades, tornando insuficientes os métodos de análise univariada, até então amplamente utilizada e consolidada.² Dessa forma, a quimiometria surgiu para tratamento e extração de informação por métodos multivariados. Esse início foi marcado por meio de uma série de artigos, no ano de 1969, sob o título “*Computerized Learning Machines Applied to Chemical Problems*”.³⁻⁶

Já o termo quimiometria (do inglês, *chemometrics*) foi criado em 1971, a partir da combinação de duas palavras suecas *kemo* (Química) e *metri* (medida), assim como sua versão em inglês, *chemo-* + *-metrics*, ambas pelo professor sueco Svante Wold.⁷ Desde então, a quimiometria tem ganhado cada vez mais espaço na discussão científica e aplicabilidade numa ampla dimensão de estudos, principalmente dentro da Química Analítica. Diversas metodologias têm sido desenvolvidas dentro das três vertentes da quimiometria, planejamento e otimização de experimentos, calibração multivariada e reconhecimento de padrões. Esta última esteve presente nos primeiros trabalhos publicados por Svante Wold, incluindo-se o método de classificação desenvolvido por ele e publicado em 1976,⁸ *Soft Independent Modeling of Class Analogy* (SIMCA), amplamente adaptado e ainda empregado. Com isso, novas abordagens são rotineiramente apresentadas frente aos grandes desafios na resolução dos problemas químicos, como de extrair com eficiência as informações químicas contidas em matrizes, cujas complexidades são muitas vezes refletidas nos dados espectroscópicos.

Neste contexto, houve um desenvolvimento simultâneo em que de um lado novos instrumentos analíticos produziam cada vez mais dados, exigindo cada vez mais métodos de análises de dados inovadores e eficazes. Os instrumentos analíticos modernos, como um equipamento de espectroscopia na região do infravermelho com transformada de Fourier, não produzem a informação com o significado químico de forma imediata. Para que essa informação traga contribuições do ponto de vista de

fenômenos químicos ou físicos, é necessária transformação e interpretação adequados. Com isso, de outro lado, houve o aumento da capacidade de computadores pessoais, permitindo cálculos mais intensivos, e o desenvolvimento de métodos de reconhecimento de padrões, elaborados e distribuídos pelos químicos para resolver problemas químicos, dando à quimiometria um papel significativo.⁹⁻¹¹

Dessa forma, a quimiometria foi avançando como uma área multidisciplinar em paralelo aos avanços da instrumentação analítica e da capacidade computacional. Ao longo dos anos essa área tem se expandido à solução de problemas de diferentes áreas com diversas abordagens, como planejamento de experimentos, calibração multivariada, reconhecimento de padrões e resolução de misturas. Ao mesmo tempo em que se avançou na complexidade e dimensão dos dados gerados pela Química Analítica moderna, como as técnicas espectroscópicas com transformada de Fourier, com produção de dados colineares, ou as técnicas hífenadas, como a cromatografia gasosa bidimensional, equipamentos multiespectrais e de imagem, entre outros, com produção de dados de altas dimensões, a quimiometria vem desenvolvendo e aprimorando métodos que atendam a essas novas demandas. São exemplos desses avanços o desenvolvimento de métodos de resolução de curvas como resolução multivariada de curvas com mínimos quadrados alternantes (MCR-ALS)¹² e o método para decomposição de dados de ordem superior PARAFAC.¹³ Outro exemplo, dentre os quais iremos abordar neste trabalho, são as adaptações de métodos de aprendizagem de máquina para solucionar problemas químicos, como a floresta aleatória (RF, do inglês *random forest*), proposta originalmente em um periódico especializado de aprendizagem de máquina, por Leo Breiman, em 2001.¹⁴ O autor demonstrou o método em diversos dados, mas não haviam dados de origem espectral. Entretanto, RF tem sido adaptada a diversos problemas do ponto de vista químico a partir de dados produzidos pelos instrumentos analíticos modernos.^{9,15-21}

Diante deste contexto, idealizamos, a princípio, um projeto de pesquisa para novas abordagens de aprendizagem de máquinas na interpretação de dados espectroscópicos no estudo qualitativo e quantitativo de matrizes complexas como petróleo e seus derivados. Entretanto, diante do desafio atual frente ao contexto da pandemia causada pelo vírus SARS-CoV-2, causador da infecção síndrome respiratória aguda denominada COVID-19, adicionamos neste trabalho uma abordagem com aplicação de aprendizagem de máquina para análise multivariada de

dados espectroscópicos (MIR) de amostras de biofluidos de pessoas infectadas e não infectadas pela COVID-19. Assim como neste estudo, diferentes métodos de quimiometria foram direcionados a um campo de estudo conhecido como bioespectroscopia, que compreende o estudo de amostras biológicas por meio da espectroscopia vibracional, principalmente na busca por uma proposta de diagnóstico menos invasivo, eficiente e rápido, como uma alternativa para triagem de pacientes com suspeita de infecção.^{22–28}

Entre os desafios desse campo de estudo pode-se citar a alta gama de variabilidade de vírus e outros microrganismos que circulam o organismo humano, além de toda a especificidade de cada indivíduo, tornando as amostras biológicas uma das mais complexas. Frente a estes desafios os métodos multivariados associados à espectroscopia molecular de biofluidos apresentam uma vantagem de primeira ordem, que é a capacidade de modelar na presença de interferentes, desde que estes estejam em todas as amostras de treinamento dos modelos.²⁹ Dessa forma, são empregados nessa abordagem métodos quimiométricos lineares e não-lineares, além de métodos de seleção de variáveis.^{23,30–34} Entretanto, nestes estudos também ocorre a dificuldade de se adquirir um conjunto amostral adequado, dificultando a obtenção de conclusões significativas. Neste caso, métodos do tipo *ensemble* podem ser uma importante alternativa e não são amplamente utilizados, de acordo com nosso conhecimento. *Ensemble* são métodos múltiplos, com combinação de resultados para resolução de um mesmo problema.³⁵ Entre os métodos do tipo *ensemble*, floresta aleatória tem alta potencialidade nos estudos no campo da bioespectroscopia.^{14,21}

RF é um método *ensemble* que combina árvores de regressão, ou de classificação, a partir de subamostragem do conjunto de treinamento e de variáveis a cada árvore.^{14,15} Sua utilização é mais ampla em reconhecimento de padrões. Na bioespectroscopia tem sido pouco aproveitada, embora os poucos trabalhos existentes tenham mostrado resultados promissores para classificação e até mesmo como forma de análise não-supervisionada.^{19–21,36,37} Seu pouco aproveitamento tem sido justificado pela dificuldade de interpretabilidade e complexidade do método, muitas vezes abordado como caixa preta ou cinza, um termo geralmente adotado para expressar um método computacional em que não se conhece bem as etapas do processamento, apenas os dados de entrada e a saída do modelo. Além disso

também não há ampla disponibilidade do algoritmo em softwares comerciais no campo das análises químicas.²¹

Dessa forma, esta tese apresenta abordagens relacionadas à compreensão da floresta aleatória em diferentes problemas de Química Analítica.

1.1. Objetivos

Objetivo geral: adaptar e aplicar métodos de aprendizagem de máquina para interpretação de dados espectroscópicos na solução de problemas em diferentes áreas da Química.

Objetivos específicos:

- 1 Aplicar método de seleção de variáveis e modelos de reconhecimento de padrões na identificação de adulteração de diesel a partir de dados espectroscópicos;
- 2 Desenvolver abordagem de aprendizagem de máquina para classificação de dados de bioespectroscopia a partir de amostras de saliva para diagnóstico de COVID-19;
- 3 Propor nova abordagem de reconhecimento de padrões por meio de aprendizagem de máquina para reconhecimento não supervisionado de padrões, seleção de variáveis e classificação, a partir de dados espectroscópicos no estudo de diagnóstico para COVID-19;

1.2. Estrutura da tese

Nesta tese são apresentadas abordagens para resolução dos problemas descritos nos objetivos específicos, com ênfase no método floresta aleatória (RF) para classificação e na forma de análise não-supervisionada (URF) de reconhecimento de padrões.

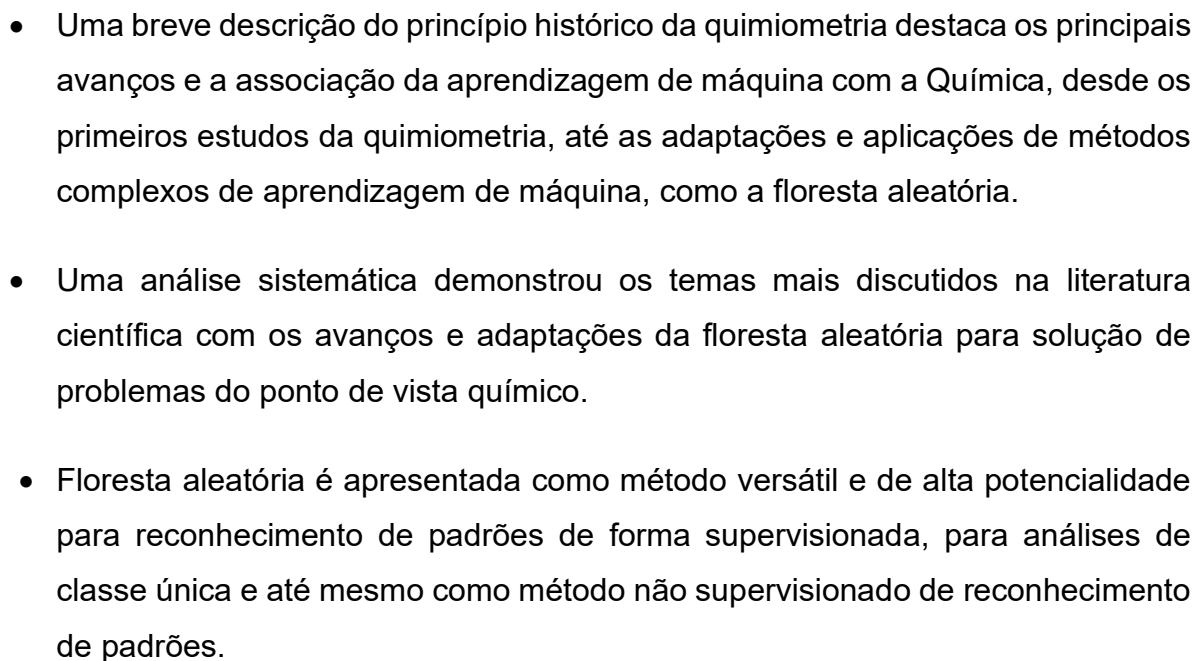
No Capítulo 2 é apresentado um estado da arte, por meio de análise sistemática pelo Bibliometrix³⁸ e escala pelo método Ordinato³⁹ a respeito da utilização de floresta aleatória na interpretação de espectroscopia em abordagens químicas, bem como a fundamentação teórica do método floresta aleatória.

No Capítulo 3 é apresentada uma proposta de aplicação da floresta aleatória não-supervisionada na triagem de adulteração de diesel combustível, por meio de um estudo de caso de suspeita de adulteração.

No Capítulo 4 é apresentada uma abordagem de reconhecimento de padrões com uso da floresta aleatória não supervisionada na identificação de estruturas em dados de bioespectroscopia a partir de amostras de saliva. Ademais, é apresentada uma classificação por consenso, com a combinação de resultados de diferentes modelos de classificação para encontrar uma resposta final com maior robustez.

No capítulo 5 é apresentada uma nova abordagem de uso da URF para reconhecimento de padrões não supervisionado, seleção de variáveis e classificação. O método foi proposto para extrair informação de espectroscopia MIR de biofluidos em *swab* nasofaríngeo, contribuindo na busca por métodos alternativos de diagnóstico de COVID-19.

No capítulo 6 são apresentadas as conclusões da tese. Por fim, nos Anexos, são apresentados os materiais suplementares dos respectivos artigos científicos e a produção científica durante o período de doutorado.



2. FLORESTA ALEATÓRIA APLICADA A ESPECTROMETRIA NA REGIÃO DO INFRAVERMELHO

2.1. Princípio histórico da Quimiometria

Pode-se afirmar que a abordagem de estatística aplicada em problemas com informações químicas tenha se iniciado décadas antes do que se costuma descrever como o princípio histórico da Quimiometria. Esta área da Química surgiu do desafio de analisar-se dados químicos de natureza multivariada, na década de 70.^{1,40,41} Entretanto, alguns métodos estatísticos empregados nesta área, como a análise de componentes principais (PCA), que é amplamente aplicada para reconhecimento de padrões, tem sua origem estatística há décadas da quimiometria. A PCA foi introduzida em 1901, por Karl Pearson. Este método foi apresentado em um artigo intitulado "*On lines and planes of closest fit to systems of points in space*". Mas sua primeira aplicação oficial ocorreu somente em 1933, por Harold Hottelling, revolucionando a análise multivariada de dados na psicologia.⁴²⁻⁴⁴ Em 1970, um artigo que demonstrou a implementação da PCA como uma expansão para seleção de características e ordenamento de dados foi utilizado como fonte para os primeiros quimiometristas.^{45,46} Outro método clássico para a quimiometria, análise discriminante linear (LDA), foi introduzida por Fisher, em 1936.⁴⁷

Em um dos primeiros artigos com a abordagem da estatística aplicada a um problema do ponto de vista químico, Fisher e Mackenzie (1923)⁴⁸ utilizaram cálculos multivariados em estudo de variedades de batatas. Os autores criticaram trabalhos anteriores que não levavam em consideração a resposta das variedades de batatas aos diferentes tratamentos com adubos. O artigo tinha como objetivo principal chamar à atenção para uma análise do tipo de dado adequado para fornecer uma resposta, assim como a elucidação do tratamento estatístico para obtenção dessa resposta. Um dos pontos fortes do trabalho foram as elucidações de fórmulas matemáticas, propostas pelos autores para aumentar a variância dos dados e distinguir entre o efeito da variedade de batatas e dos fertilizantes. Destacaram ainda que a solução dada não seria necessariamente aplicada a outros casos, impulsionando uma visão heurística da estatística aplicada à química.

Cabe aqui ressaltar que a aplicação citada acima não contava com informação química gerada por algum tipo de medida instrumental. Mas a partir dos avanços

tecnológicos na área instrumental, uma importante relação entre a Química e a estatística multivariada tem sido presente desde a década de 70. Pois, nesta época a computação científica tornou-se de certa forma acessível, permitindo conectividade de equipamentos a computadores.¹ Esse avanço nos métodos de análise propiciou um aumento considerável na geração e na complexidade dos dados analíticos. Dessa forma então, surgiu o desafio de extrair e interpretar a informação química a partir desses novos dados, produzidos pela moderna instrumentação, desde então interfaciada a computadores. Para interpretação de espectros produzidos pelos instrumentos ópticos como infravermelho, por exemplo, necessitava-se de grande biblioteca de dados espectroscópicos e espaço para armazenar estes dados.^{1,6}

Os métodos univariados até então amplamente utilizados não seriam suficientes, mediante a demanda de informação sendo gerada e dos interferentes inerentes à complexidade dessa informação.² Neste cenário é que surgiu a Quimiometria, utilizando, a princípio, as técnicas de estatística aplicada para solucionar problemas químicos. Seu início histórico foi marcado por uma série de artigos, no ano de 1969, sob o título "*Computerized Learning Machines Applied to Chemical Problems*."^{3-6,49}

No primeiro da série, Jurs, Kowalski e Isenhour³ descreveram a utilização de aprendizado de máquina com classificação binária na identificação de fórmulas moleculares. Eram dados de 346 espectros de espectrometria de massas de baixa resolução, originados de Projeto de Pesquisa do Instituto Americano de Petróleo (API, do inglês, *American Petroleum Institute*). Os dados foram dispostos em cartões magnéticos próprios para utilização nos computadores com linguagem de programação Fortran (compilador Fortran IV). Os autores determinaram fórmulas moleculares com confiabilidade de 98% e destacaram o potencial dos computadores e de métodos de aprendizado de máquina na interpretação de dados químicos.

Em outro artigo, Kowalski, Jurs e Isenhour⁶ descreveram a implementação de aprendizado de máquina para a classificação entre as classes químicas de 4500 espectros de infravermelho de compostos-padrão, voltando-se à interpretação dos espectros. A partir dessa série de artigos, o desenvolvimento e a aplicação de métodos multivariados passaram a ser objeto de estudo na quimiometria, para solução de problemas em diversas áreas da Química.

Em 1971, Svante Wold criou o termo quimiometria (do inglês *chemometrics*) numa versão sueca, pela combinação das palavras *kemo* (Química) e *metri* (medida), assim como sua versão em inglês, *chemo-* + *-metrics*.^{1,2} Neste mesmo período dois artigos tiveram impacto na ampliação dos métodos de reconhecimento de padrões, o de Kowalski e Bender,⁵⁰ que é considerado o primeiro artigo da área, e de Lawstron e Sylvestre.⁵¹ Kowalski e Bender⁵⁰ apresentaram abordagens de utilização de método não-paramétrico de reconhecimento de padrões (que não faz suposição sobre a distribuição estatística subjacente dos dados).

No ano seguinte, Svante Wold formou o primeiro grupo de pesquisa em quimiometria e publicou o primeiro artigo usando este termo. No ano de 1974 os professores Svante Wold e Bruce Kowalski fundaram a Sociedade Internacional de Quimiometria, com o princípio de fortificar o elo entre matemática aplicada, estatística e aplicações na área da química. Este grupo conceituou que “a arte de extrair informação química relevante de dados produzidos em experimentos químicos é chamada de quimiometria em analogia à biometria e econometria”.^{2,52} No primeiro artigo com o termo quimiometria no título, Kowalski referiu-se à quimiometria como um “conjunto de métodos apropriados para a obtenção de informações significativas, sob o ponto de vista químico”.^{1,2} E no ano de 1976, ocorreu o primeiro simpósio de quimiometria, “*Chemometrics: Theory and Applications*”, organizado pela divisão de computação em Química da Sociedade Americana de Química, originando o primeiro livro da área, publicado em 1977.

Concretizando a quimiometria como uma nova área da Química, na década de 80 foram fundados os dois periódicos voltados exclusivamente para esta área: *Chemometrics and Intelligent Laboratory Systems* (Chemolab, 1986 – Elsevier) e *Journal of Chemometrics* (1987 – Wiley).^{1,2,52}

O Brasil também já havia sido inserido nos estudos de desenvolvimento da quimiometria nesta mesma época. Em 1978 fora enviado pelos correios ao Instituto de Química da Universidade de Campinas, UNICAMP, aos cuidados de Prof. Dr. Roy Edward Bruns, uma fita magnética de 800 m de comprimento, com o programa ARTHUR, que foi desenvolvido pelo grupo do professor Kowalski. O programa continha algoritmos no compilador Fortran IV para reconhecimento de padrão e cálculos estatísticos multivariados. Na Figura 2.1 Prof. Dr. Roy Edward Bruns, em 2020 no workshop de quimiometria,⁴⁶ mostra a fita magnética recebida em 1978.

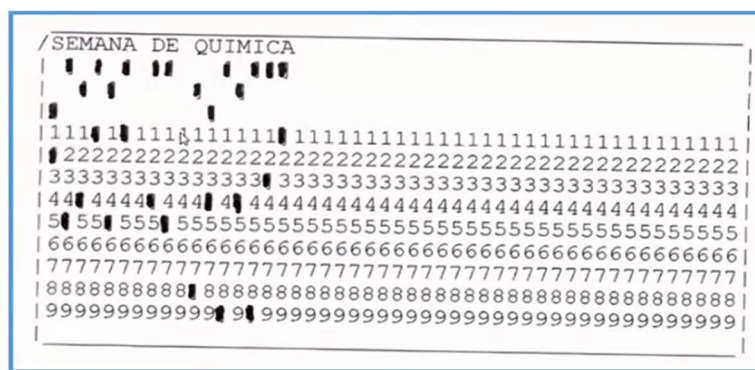


Figura 2.1 Prof. Dr. Roy Edward Bruns demonstrando fita magnética, em web conferência para mesa redonda em XI Workshop de Quimiometria.⁴⁶

Fonte: Captura de tela extraída de web conferência em XI Workshop de Quimiometria, 2020,⁴⁶ sob autorização de Prof. Dr. Roy Edward Bruns, em 26/01/2022.

Os dados eram disponibilizados em cartões perfurados (Figura 2.2, A) que eram lidos e processados no computador. Este último compreendia um equipamento PDP-10 (Figura 2.2, B), um dos mais avançados da época entre as universidades brasileiras. Era um equipamento multiusuário da UNICAMP, com 1MB (Um megabyte) de memória RAM. Os resultados eram consultados somente por meio de lista impressa em uma impressora de linha. Em 1980, os professores Kowalski e Roy Bruns realizaram no Brasil no Instituto de Química da UNICAMP o primeiro curso intensivo de quimiometria. E, em 1981, foram defendidas as primeiras defesas de mestrado e doutorado de pesquisadores brasileiros, orientadas pelo professor Roy Bruns.^{46,53}

(A)



(B)



Figura 2.2. Cartão perfurado para análise de dados de pesquisadores do Instituto de Química (A) e computador PDP10 da marca Digital, pertencente à Universidade de Campinas - UNICAMP (B), onde eram realizadas as análises dos dados.

Fonte: captura de tela extraída de web conferência em XI Workshop de Quimiometria, 2020,⁴⁶ sob autorização de Prof. Dr. Roy Edward Bruns, em 26/01/2022.

Entre os primeiros estudos desenvolvidos sob orientação de Prof. Roy Bruns, um em especial iniciou uma ampla divulgação da calibração multivariada no Brasil e na América do Sul. Estudo este desenvolvido pelo Prof. Dr. Roney Jesus Poppi em seu curso de mestrado pela UNICAMP, em 1986. A partir desse estudo, que resultou na primeira publicação sul-americana com mínimos quadrados parciais (PLS) e regressão por componentes principais (PCR), Roney Jesus Poppi iniciou uma carreira profissional de imensa contribuição para formação de diversos grupos de pesquisa no Brasil e América do Sul. Ao longo de 30 anos, Roney Jesus Poppi fundou um grupo de pesquisa consolidado no Brasil e importante para toda a América do Sul, criou o Laboratório de Quimiometria em Química Analítica (LAQQA) da UNICAMP, orientou muitos pesquisadores dos níveis de graduação ao pós-doutorado, 290 publicações

com diversas abordagens e mais de seis mil citações nas bases de artigos científicos. Estima-se que mais de vinte de seus ex-alunos atualmente sejam coordenadores de grupos de pesquisa em Quimiometria nas universidades públicas e privadas pelo Brasil.⁵⁴ Seus principais trabalhos abordavam quimiometria com técnicas de espectroscopia molecular, sendo a maioria sobre diversas estratégias de calibração multivariada. Infelizmente, Prof. Dr. Roney Jesus Poppi faleceu em 2020, durante a pandemia causada pelo coronavírus, deixando um importante legado na formação e capacitação de pessoas na Quimiometria. Após sua morte, vários periódicos científicos lançaram chamadas para publicações ou até mesmo edições especiais em sua homenagem, como a edição de número 32 de *Brazilian Journal of Analytical Chemistry*, dedicado ao renomado professor, cuja capa está demonstrada na Figura 2.3.^{54–56}



Figura 2.3 Capa de edição dedicada ao Prof. Dr. Roney Jesus Poppi, periódico científico Brazilian Journal of Analytical Chemistry, em 2021. Fonte: BrJAC, volume 8, n.32 (2021)⁵⁵

Ao longo do desenvolvimento da Quimiometria como uma área da Química, a definição do termo quimiometria difere um pouco entre regiões, devido a diferenças linguísticas do termo, entre outros fatores⁵². No Brasil, a professora Dr.^a Márcia Miguel Castro Ferreira conceitua o termo quimiometria como “administração e processamento de informações de natureza química, com utilização de ferramentas matemáticas e estatísticas, aplicada a diversas áreas da química.”¹¹ Em 2016, a União Internacional de Química pura e aplicada (IUPAC) definiu quimiometria como “a ciência de relacionar medidas feitas em um sistema ou processo químico ao estado do sistema (expressos como valores de propriedades macroscópicas) via aplicação de métodos matemáticos ou estatísticos”.⁵⁷

Pode-se observar que a Quimiometria avançou paralelamente com os avanços computacionais e instrumentais. Atualmente é possível ver sua aplicação em processos industriais, em softwares acoplados a equipamentos, entre outros. Embora seja uma área multidisciplinar dentro da Química Analítica, é um campo de estudo abrangente, que continua expandindo de acordo com o aumento da demanda pela análise de dados cada vez mais complexos. Isso requer cada vez mais estudantes interessados nas habilidades para desenvolver métodos dentro deste campo de estudos, integrando-se os avanços nas áreas de estatística multivariada, quimiometria e aprendizagem de máquina, numa visão heurística.

2.2. Reconhecimento de padrões e Aprendizagem de máquinas

Desde seu princípio histórico, diversas metodologias têm sido desenvolvidas dentro das três vertentes da Quimiometria, planejamento e otimização de experimentos, calibração multivariada e reconhecimento de padrões. Este último abrange o estudo realizado neste trabalho. Entretanto, os termos “reconhecimento de padrões”, “aprendizagem de máquinas”, “mineração de dados” e “classificação”, têm sido sobrepostos em algumas aplicações.^{58,59}

Reconhecimento de padrões é definido nos primeiros artigos como toda abordagem computacional usada para determinar padrões ou relações entre objetos. Entretanto, não há uma definição universal, mas uma visão em comum nas definições modernas envolvendo principalmente classificação ou atribuição de objetos a grupos ou classes. Estes métodos de classificação podem ser divididos em dois tipos: supervisionado, no qual os objetos são divididos em grupos de acordo com suas

categorias pré-definidas, por meio de um conjunto de treinamento; e, não-supervisionados, com divisão dos dados no espaço em grupos sem qualquer conjunto de treinamento pré-definido.^{9,58}

Outra divisão possível para os métodos de reconhecimento de padrões são os métodos puramente estatísticos ou de aprendizado de máquina, o que não está claramente distinto na literatura, atualmente. Os métodos considerados principalmente estatísticos podem ser compreendidos como os que utilizam uma equação linear discriminante para separar o conjunto amostral em duas categorias, por exemplo. Já os métodos que envolvem aprendizado de máquina, são os que requerem maior recurso computacional, com bases estatísticas nos algoritmos, principalmente estatística bayesiana, mas com recursos computacionais mais intensos de aprendizagem de máquina para extração de fatores e identificação de padrões.^{9,58}

Assim como o termo quimiometria, aprendizagem de máquinas também possui uma diversidade de definições. Jurs Kowalski e Isenhour³ definem como classificadores de padrões treináveis, ou sistemas inteligentes que recebem estímulos e aprendem a selecionar uma resposta correta. Esse aprendizado se dá por meio de *feedback* iterativo durante processo de decisão, de forma a maximizar a probabilidade de respostas desejadas.

Os algoritmos de aprendizagem de máquinas utilizam métodos computacionais para “aprender” informação diretamente dos dados sem depender de uma equação pré-definida como modelo. Os algoritmos melhoram adaptativamente o seu desempenho à medida que o número de amostras disponíveis para aprendizagem aumenta.^{59,60} A subdivisão entre métodos supervisionados e não-supervisionados é a mesma descrita anteriormente. As etapas comuns aos métodos supervisionados são demonstradas no esquema da Figura 2.4.

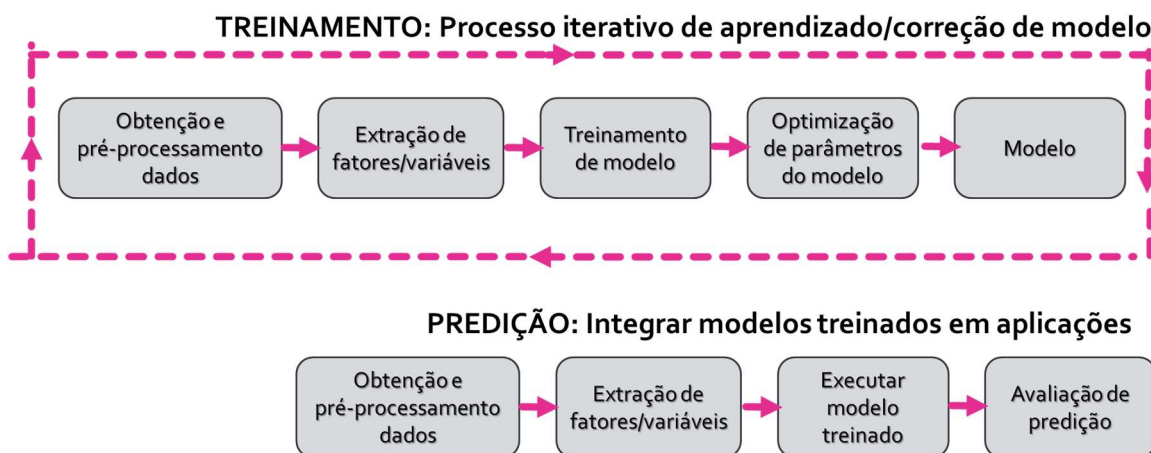


Figura 2.4 Esquema geral de etapas iterativas para construção e aplicação de modelos de aprendizagem de máquina a dados químicos.

Fonte: compilado pela autora, adaptado de MATHWORKS, e-book, 2021.⁶⁰

Apesar de afirmar não haver uma distinção clara a respeito da sobreposição dos termos, para Brereton,⁵⁸ um método como LDA pode ser considerado como de base estatística, não se enquadrando em aprendizagem de máquinas. Já como método de aprendizagem de máquina são exemplos: máquina de vetor de suporte (SVM, do inglês *support vector machine*),⁶¹ redes neurais artificiais (ANN, do inglês *artificial neural networks*)⁶² e floresta aleatória (RF)¹⁴ entre outros. Mas métodos com função discriminante também são considerados como aprendizagem de máquina em e-book distribuído por desenvolvedores.^{60,63} e em alguns artigos,^{64,65} mostrando que não há um consenso a respeito dessa divisão. Brereton^{1,58} relata que os primeiros artigos mostravam desenvolvimento de aprendizado de máquina, mas ainda no princípio histórico os próprios autores modificaram o foco para reconhecimento de padrões que é mais abrangente.

Entre os métodos atualmente aplicados a esse propósito, floresta aleatória perpassa pelas diferentes áreas da Química como um método de aprendizagem de máquina amplamente aplicado em problemas de reconhecimento de padrões.^{15–20,37,66} Este método é um *ensemble* (conjunto de modelos que compõem uma resposta final) baseado em árvores de decisão. RF foi proposta por Breiman, em 2001,¹⁴ para melhorar classificação de diversos dados a partir de amostragem aleatória e seleção de variáveis. Pode ser aplicado em problemas qualitativos ou quantitativos. Mas neste trabalho daremos enfoque às abordagens de maior relevância dentro do ponto de vista qualitativo, ou seja, de reconhecimento de padrões.

2.3. Floresta aleatória aplicada à espectrometria na região do infravermelho

Floresta aleatória é um conjunto de árvores de decisão. Uma árvore de decisão é um tipo de algoritmo de aprendizagem supervisionada, que tem sua aplicação em variáveis categóricas e contínuas. A árvore de decisão pode ser utilizada em problemas de classificação, formando uma árvore de classificação, ou de calibração multivariada, com a árvore de regressão. Apesar de a utilização das árvores de decisão serem retratadas antes da década de 90, sendo aprimoradas para problemas de regressão e classificação de dados, havia um desafio sendo debatido pelos estatísticos que era o de alcançar máxima exatidão sem causar um sobreajuste do modelo ao conjunto utilizado para treiná-lo, o que é conhecido como *overfitting*.^{14,67-70} Árvore de classificação e regressão (CART) é um algoritmo de repartição recursiva que foi inicialmente desenvolvido por Leo Breiman e outros em 1984.⁷¹ As árvores podem ser treinadas como *pruned tree* (árvore seca, podada) ou *unpruned tree* (não seca, não podada). A *unpruned tree* cresce até o tamanho máximo do conjunto de treinamento, o que pode causar o sobreajuste (*overfitting*). Isso ocorre porque conforme é crescida uma árvore, as regras de divisão ficam mais detalhadas ao ponto de ao invés de identificar regras maiores, que correspondam aos relacionamentos reais e confiáveis nos dados, a árvore passa a identificar regras menores ou específicas, que refletem apenas ruídos nos dados de treinamento. Uma árvore *unpruned* pode resultar em folhas com 100% de pureza, ou seja, 100% de exatidão na classificação de um conjunto de treinamento. Entretanto, esse resultado possui ajuste de ruído nos dados de treinamento, e não os sinais analíticos de interesse.^{14,35,67-70}

Para resolver esse desafio, Ho e colaboradores^{67,70} propuseram combinação de classificadores e árvores em subespaços selecionados aleatoriamente. Este método foi chamado de floresta de decisão, no qual foi proposto um procedimento de seleção randômica de dimensão menor a partir da dimensão original dos dados e produção de árvores diferentes.

Essa combinação de classificadores é geralmente denominada como um *ensemble*, que são métodos que treinam múltiplos classificadores e ao final combinam os resultados desses classificadores para resolver um mesmo problema. Essa ideia

de combinar modelos múltiplos tem sido usada há tempos na sociedade, mas a discussão em volta dos métodos *ensemble* aumentou significativamente a partir de 1990 em diversas áreas de pesquisa.³⁵ Essa área foi impulsionada por discussões de três abordagens: combinação de classificadores, *ensemble* de modelos fracos para gerar um modelo de maior robustez e mistura de especialistas. Sendo a combinação de classificadores a abordagem de maior relevância na área de reconhecimento de padrões.^{14,69}

A floresta aleatória (RF) é um *ensemble* proposto por Breiman, em 2001,¹⁴ a partir da fusão dos algoritmos CART (*Classification and Regression Trees*) e *bootstrapping aggregation (bagging)*.⁶⁹ O primeiro é o algoritmo de construção de árvore de classificação ou de regressão e o segundo um método de reamostragem aleatória com reposição. O *bagging* é usado na floresta aleatória para compor o conjunto de calibração (ou de treinamento), chamado *In-Bag*, e o conjunto teste, chamado *Out-of-Bag* (OOB). Breiman comparou RF principalmente com um método denominado Adaboost, que também é um *ensemble* mas sem elementos aleatórios.

A fusão dos métodos CART e Bagging adicionou uma aleatoriedade que já estava contida no *bootstrapping* para desenvolvimento de combinação de árvores de decisão.⁷² Assim, na RF cada árvore de classificação é crescida de forma independente utilizando uma amostragem aleatória. Essa foi a proposta para solucionar o problema do *overfitting*. Além disso, há outras importantes mudanças no método RF em relação aos métodos anteriores, principalmente no processo de divisão dos nós. Na RF, cada árvore é crescida com uma seleção aleatória das variáveis originais, das quais são selecionadas as variáveis preditoras, que formarão os nós de decisão. Este subconjunto de variáveis é um parâmetro de otimização, comumente denominado de *mtry*. Outro parâmetro que pode ser otimizado é o *nTrees*, que é o número total de árvores a serem construídas em determinado modelo (floresta).

Em seu artigo original, Breiman¹⁴ não orientou essa otimização, mostrando alguns experimentos com 100 árvores ao demonstrar sua proposta. Mas relatou ter escolhido 100 árvores para uma previsão confiável do conjunto de *out-of-bag*. Em outro momento, ao demonstrar a análise de variáveis importantes, utilizou 1000 árvores. Apesar de RF ser rápida em comparação com outros métodos da época, um número alto de árvores pode promover um aumento no tempo de processamento.¹⁴

Nesse contexto, Santana e colaboradores otimizaram modelos RF com dados de infravermelho médio e próximo com $nTrees = 500$.^{17,18,73} Afanador e colaboradores também relataram que, geralmente, 500 árvores são suficientes para um bom desempenho.²¹ Entretanto, Lovatti e colaboradores identificaram que um número de 1000 árvores foi necessário para processamento de dados de ressonância magnética nuclear de ^1H (RMN de ^1H) e 2000 árvores para dados de RMN de ^{13}C em amostras de petróleo bruto.^{15,16} Relataram ainda que ao adicionar métodos de redução de dimensionalidade prévia ao RF, este número de árvores reduziu para 500, para RMN de ^1H , e 1500, para RMN de ^{13}C .¹⁶ Wu e colaboradores⁷⁴ também otimizaram número de árvores em modelos de floresta aleatória, sendo necessário $nTrees = 955$ em espectros de infravermelho na região do médio (MIR) e $nTrees = 1035$ em dados de espectroscopia na região ultravioleta e visível (UV-VIS).

Outro parâmetro otimizado é o valor de $mtry$, que é o subconjunto de variáveis originais, selecionadas aleatoriamente como candidatas à construção da árvore, a cada iteração. Geralmente, os valores padrão para $mtry$ são $\sqrt{\text{variáveis}}$ para classificação, e $1/3$ das variáveis para problemas de regressão.^{75,76} Entretanto, este valor tem sido otimizado como uma razão das variáveis, a depender do tipo de dado a ser modelado, que pode variar de 1 a 99% das variáveis originais.^{15,16,18,74,76}

O que há em comum nos trabalhos citados acima é que o número de árvores escolhido deve ser o suficiente para uma estabilização na diminuição do erro de previsão do conjunto *Out-of-Bag*. De acordo com a quantidade e complexidade de amostras e de variáveis esse valor poderá mudar, mas há um ponto em que o aumento do número de árvores não produz diferença estatística significativa. Isso ocorre porque a contribuição de cada árvore passa a ser reduzida com o aumento da floresta.^{14,15,21,73,74}

Assim que proposta, RF tem sido amplamente discutida em diversas áreas, como em estudos ambientais ou de solos com uso de dados de satélites, de sensores remotos ou imagens multiespectrais.^{77–81} Suas primeiras aplicações foram relatadas com espectrometria de massas a partir de 2003.^{82–84} Além disso, a associação da RF com a espectrometria de emissão óptica com plasma induzido por laser (LIBS) também tem alcançado sucesso na classificação de matrizes alimentícias^{85–88} e outras matrizes.^{89–92}

Entretanto, a implementação da RF à interpretação de dados espectroscópicos, que é o foco principal deste trabalho, foi realizada somente 6 anos após sua introdução. No primeiro trabalho nesse sentido, Menze, Petrich e Hamprecht⁹³ utilizaram RF para selecionar regiões espectrais em espectros MIR de 641 amostras de soro bovino para estudo de diagnóstico de doença veterinária. Além de selecionar as variáveis, a RF foi utilizada como modelo de classificação não-linear em um sistema hierárquico de classificação no qual combinaram em diferentes etapas classificadores lineares e não-linear. Atualmente a maior parte das discussões associando RF a dados espectroscópicos compreende estudos qualitativos e quantitativos com amostras de solo e alimentos.^{17,18,73,77,85,94–96} Além de aplicações em diversas matrizes e diferentes tipos de equipamentos, como alimentos em sensores eletrônicos,⁹⁷ alimentos com imagem multiespectral,^{97,98} alimentos e outras matrizes com equipamentos espectroscópicos,^{15,77,99,100} estudos bioquímicos, metabolômica e genômica,^{82,101} microarranjo para perfil genômico,^{102,103} atividade biológica de compostos moleculares,⁷⁶ entre outros. Dessa forma, uma análise ampla de todas as citações dos termos “*random forest*” e “*spectroscop*” pode resultar em mais de 3.000 artigos nas bases *Web of Science* e *Scopus*.

Mas uma análise sistemática^{38,39} em busca dos termos citados acima em títulos, palavras-chave, resumos e tópicos, nos últimos 5 anos, resultou em uma série de aproximadamente 800 artigos indexados nas bases *Web of Science* e *Scopus*. Destes, cerca de 450 são soluções para problemas qualitativos. Entre os estudos encontrados pela análise sistemática, como previsto a maior parte contém os termos solo e alimentos, mas também uma boa quantidade de estudos de diagnóstico de doenças, como pode ser observado na Tabela 2.1 que traz alguns termos-chave encontrados dentro do resultado da análise sistemática.

Tabela 2.1 Ocorrência de termos-chave a partir de uma pesquisa qualitativa nos resultados de bibliometria com termos-chave “Random forest” AND “Spectroscop”

Termo pesquisado em resultado de análise sistemática	Número de documentos
Solo	246
Alimentos	73
Combustível	5
Petróleo bruto	7
COVID	5
Diagnóstico	104
Câncer	75

Fonte: compilado pela autora a partir de bibliometria realizada, 2023.

Entre os estudos indexados com aplicação da RF à interpretação de dados espectroscópicos são encontradas várias técnicas, como espectroscopia de reflectância na região do infravermelho próximo e do visível (VIS-NIR), infravermelho na região do médio (MIR) e do próximo (NIR), espectroscopia Raman, entre outras. Uma amostragem de algumas das técnicas espectroscópicas empregadas nos estudos associadas com RF, com base nos 100 artigos mais relevantes^{38,39} dos últimos 5 anos, está apresentada na Tabela 2.2. Também é possível observar a aplicação da RF com a combinação de dados de técnicas diferentes. A fusão de dados de nível baixo, que é um método que realiza a junção dos dados espectroscópicos no seu formato original, aumenta a quantidade de informação. A RF pode ser um método aplicado para modelar este novo conjunto de dados, pelas características de um modelo no formato *ensemble* e subamostragem.

Tabela 2.2 Amostragem de estudos associando técnicas espectroscópicas a modelos RF.

Técnica	Tipo de abordagem	Número de documentos	Ref.
UV-VIS-NIR	Contaminação de alimento/agricultura	02	104,105
VIS-FIR	Classificação de espécies de plantas	01	106
VIS-NIR	Análise qualitativa na adulteração de alimentos	01	17
	Análise quantitativa na adulteração de alimentos	01	100
	Estimativa de teor de carbono orgânico em e matéria orgânica em solo	04	107–110
	Estimativa de propriedades físico-químicas de solo	04	18,94,111,112
	Estimativa de contaminação de solo por metais pesados	04	113–116
	Estimativa de contaminação de solo por hidrocarbonetos de petróleo	02	117,118
	Caracterização de plantas	01	119
	Estimativa de propriedades físico-químicas de vidros e silicatos	01	120
	Origem geográfica de alimentos e teor de contaminante em alimentos	02	95,96
MIR	Origem geográfica, adulteração ou contaminação de alimentos	03	17,121,122
	Classificação de biofluido para detecção de câncer	01	123
MIR-NIR	Origem geográfica de alimentos e fármacos fitoterápicos	03	124–126
MIR-UV/VIS	Origem geográfica de fármaco fitoterápico	01	74
	Caracterização de alimentos	03	127–129
NIR	Estimativa de propriedades físico-químicas de biodiesel	01	130
Continua...			

Continuação			
Técnica	Tipo de abordagem	Número de documentos	Ref.
NIR-PXRF	Identificação de subgrupos de solos	01	131
PXRF	Identificação de perfil de solos	01	132
	Estimativa de propriedades físico-químicas de solo	05	84,112,133–135
Raman	Identificação de perfil de microplásticos em espécies marinhas	01	121
	Identificação de tipos de leites	01	136
	Identificação de perfil metabolômico de meio de cultura de fertilização <i>in vitro</i>	01	137
RMN ¹ H	Classificação de soro para estudo metabolômico	02	138,139
	Identificação de tipos de vinhos	01	140
SERS-Raman	Classificação de biofluidos / detecção doenças	04	141–144

UV-VIS-NIR: espectroscopia de reflectância na região do ultravioleta (100 – 400 nm), visível (400 – 780 nm) e infravermelho próximo (14.290 – 4.000 cm⁻¹); VIS-FIR: espectroscopia de reflectância na região do visível (400 – 780 nm) e infravermelho distante (10 – 200 cm⁻¹); VIS-NIR: espectroscopia de reflectância na região do visível (400 – 780 nm) e o infravermelho próximo (14.290 – 4.000 cm⁻¹); XAS: Espectroscopia de absorção de raios X; MIR: Espectroscopia infravermelho na região do médio (4.000 – 200 cm⁻¹); NIR: espectroscopia infravermelho na região do próximo (14.290 – 4.000 cm⁻¹); UV: espectroscopia na região do ultravioleta (100 – 400 nm); PXRF: fluorescência de raio X portátil; RMN ¹H: ressonância magnética nuclear de ¹H; SERS-Raman: espalhamento Raman intensificado por superfície. Fonte: Adaptado de análise bibliométrica pela técnica de bibliometrix.³⁸ 15/01/2023.

Em termos de número de produção científica, em 2022 observou-se um aumento significativo de produção de artigos científicos, de acordo com os termos da análise sistemática realizada, seguindo uma tendência de crescimento nos últimos cinco anos (Figura 2.5).

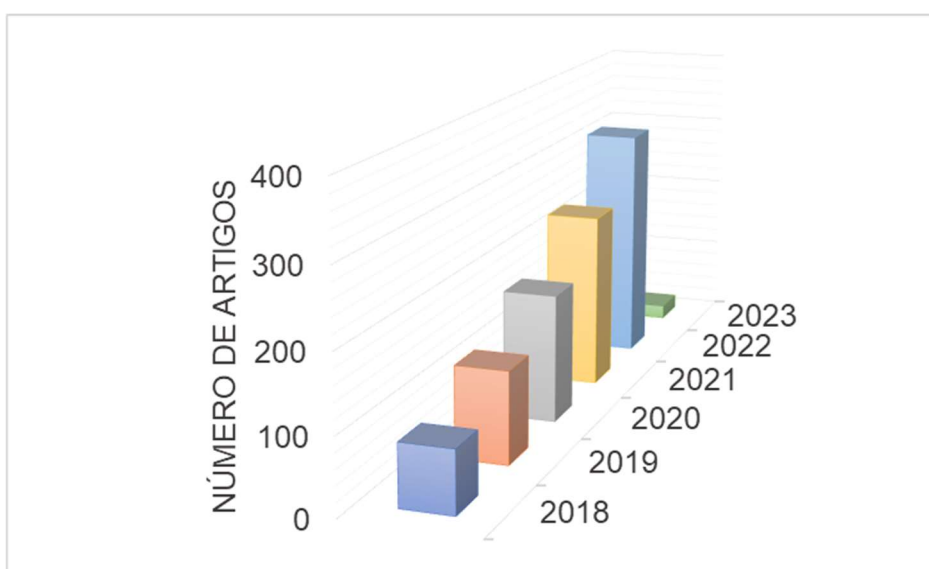
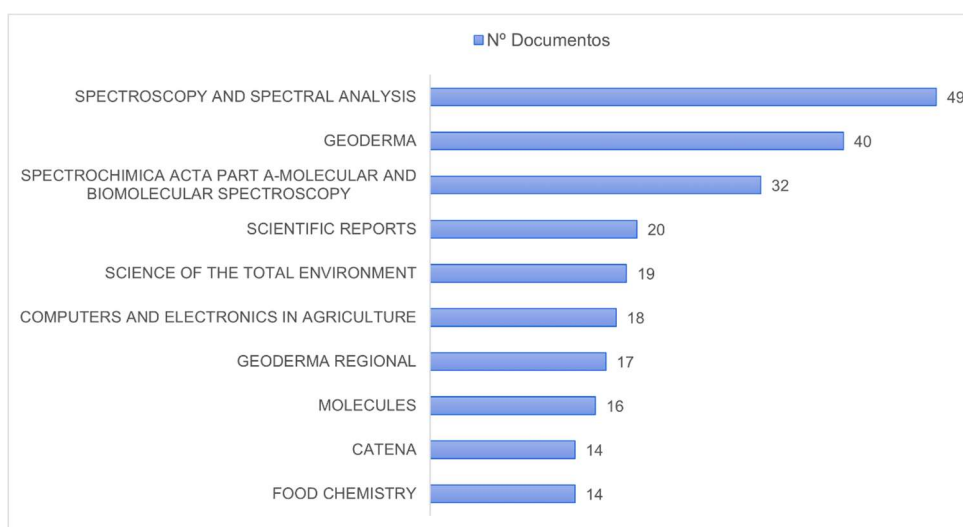


Figura 2.5 Número total de artigos publicados, por ano, a partir de análise bibliométrica nas bases *Web of Science* e *Scopus* sobre a associação de RF e espectroscopia nos anos de 2018 a 2023 (15/01/2023).

Fonte: Adaptado de análise bibliométrica pela técnica de bibliometrix.³⁸ 15/01/2023.

As diversas produções retêm um maior número em periódicos voltados à aplicação do método nas diferentes matrizes e desafios (Figura 2.7 A), acompanhando uma tendência citada anteriormente neste estudo. Em termos de relevância pelo número de citações, ao se observar os dez periódicos mais relevantes inclui-se um dos periódicos voltados ao desenvolvimento e aplicações de métodos quimiométricos, *Chemolab*, da *Elsevier*, como pode ser observado na Figura 2.7 B.

A)



B)

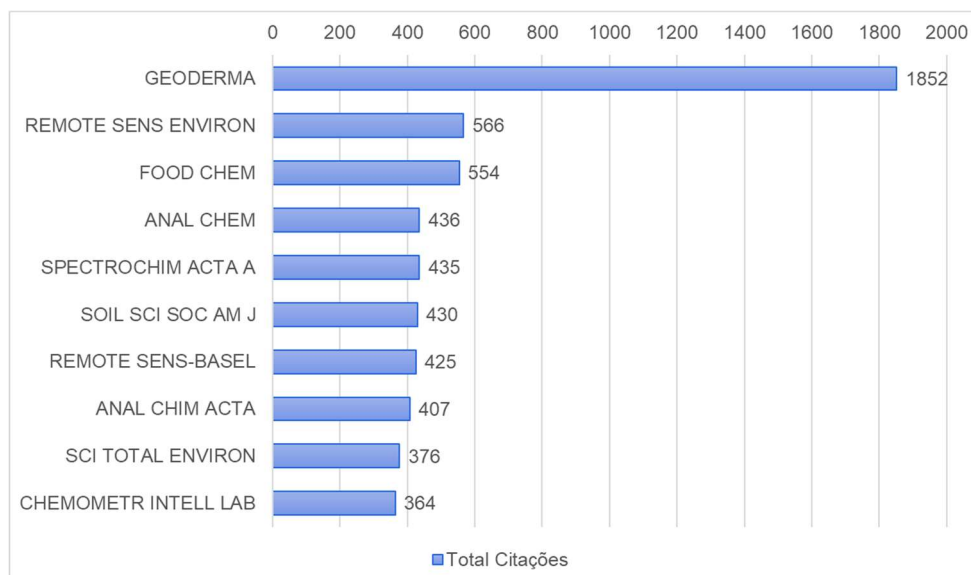


Figura 2.6 Dez fontes mais relevantes em número de documentos (A) e total de citações (B), a partir de análise bibliométrica nas bases *Web of Science* e *Scopus* sobre a associação de RF e espectroscopia nos anos de 2018 a 2023 (15/01/2023).

Fonte: Adaptado de análise bibliométrica pela técnica de bibliometrix.³⁸ 15/01/2023.

Entre os países de maior produção encontram-se China, Estados Unidos e Brasil (Figura 2.7, A). Entretanto, em termos de colaboração entre os países o Brasil faz poucas contribuições com os maiores produtores, sendo concentrada a contribuição entre China e Estados Unidos (Figura 2.7, B).

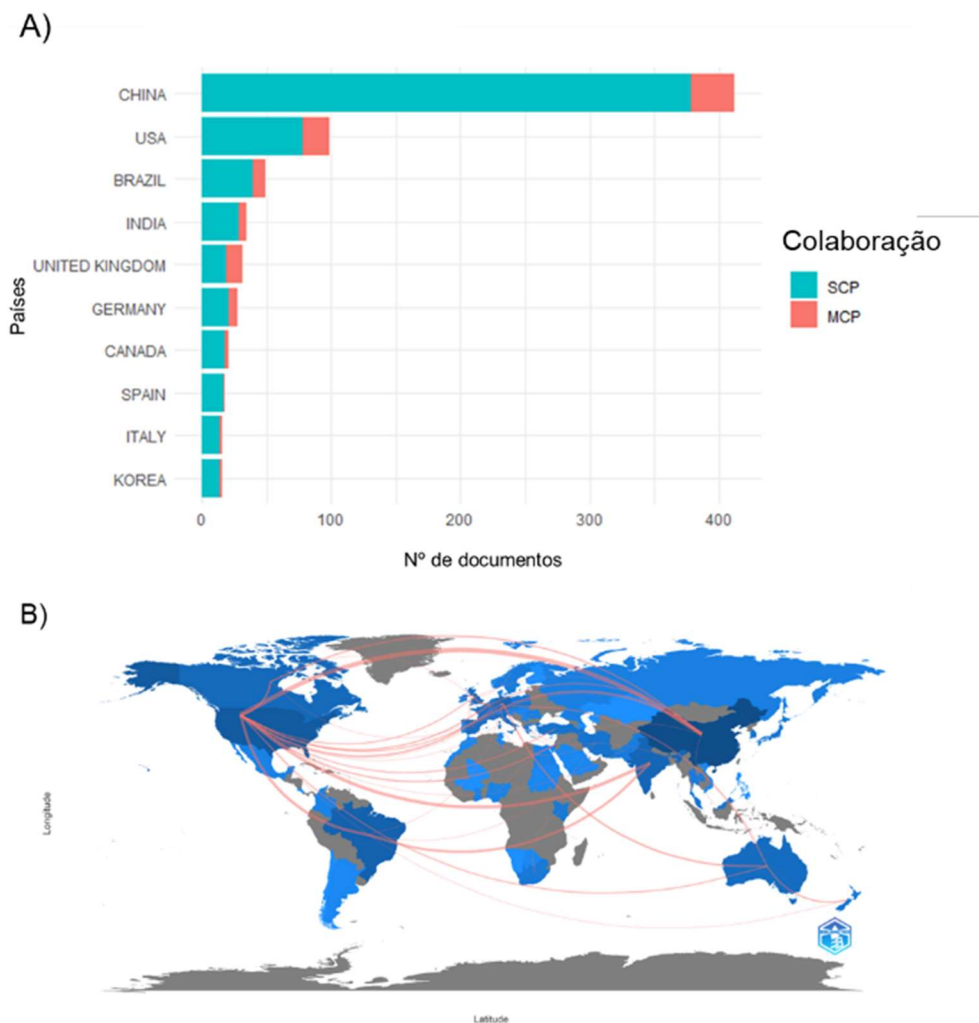


Figura 2.7 Produção e contribuições entre países com associação dos termos floresta aleatória e espectroscopia no período de 2018 (01/01/2018) a 2023 (15/01/2023). (A) Dez países mais produtivos (SCP: produções sem contribuições entre países; MCP: produções com contribuição entre países); contribuição e com contribuição; (B) Mapa de colaboração entre os países.
Fonte: Adaptado de análise bibliométrica pela técnica de bibliometrix.³⁸ 15/01/2022.

Em uma das aplicações mais relevantes nos últimos 5 anos,^{38,39} Li, Zhang e Wang ¹²⁴ realizaram fusão de dados de espectroscopia infravermelho na região do próximo (NIR) ($10.000 - 4.000 \text{ cm}^{-1}$) e de MIR ($4.000 - 400 \text{ cm}^{-1}$) de uma planta medicinal de muita exploração, devido sua utilização pela medicina chinesa. Com dados de cinco maiores produtores, objetivavam autenticação de origem geográfica.

Com os dados processados com padronização da variação de sinais (SNV) e segunda derivada, utilizaram RF com número de variáveis padrão e 2000 árvores para selecionar variáveis pelo índice de importância a partir do índice de Gini. Dados com variáveis selecionadas e fusão de nível médio foram classificados pela RF com 95 a 98% de exatidão. Já a fusão de nível alto abrangendo modelo de SVM e RF obtiveram classificação com 98 a 100% de exatidão.

Outro tipo de problema de autenticação foi abordado por Santana e colaboradores ⁷³ a partir de amostras de óleo de andiroba. Devido ao seu valor comercial e demanda, este tipo de produto tem sido adulterado com misturas de diferentes óleos vegetais, de propriedades físico-químicas semelhantes. Os autores simularam adulteração com óleo de milho e óleo de soja em diferentes proporções 192 amostras de óleo de andiroba. Os modelos de classificação foram propostos com RF a partir de dados de espectroscopia MIR (número de onda: 3100 – 680 cm⁻¹) com reflectância total atenuada horizontal (HATR), tendo 40 amostras autênticas e 192 adulteradas. O modelo com 500 árvores e número padrão de variáveis, pré-processadas pela primeira derivada, como dados de entrada, obteve 100% de exatidão no conjunto de teste.

RF também tem sido bem-sucedida em outras aplicações de classificação pela origem geográfica a partir de dados espectroscópicos de plantas medicinais da Ásia ^{74,124,126}. Wu, Zhang e Wang ⁷⁴ analisaram 161 amostras de uma planta medicinal asiática (*Paris polyphylla var. yunnanensis*) de seis diferentes regiões. As amostras foram limpas, secas a 50 °C e moídas para análise em pó. Foram adquiridos espectros de MIR (número de onda de 4000 - 650 cm⁻¹) e UV-VIS (comprimento de onda de 200 - 500 nm). Com os dados pré-processados com SNV e suavização Savitzky-Golay pela segunda derivada (MIR) e primeira derivada (UV-VIS), a fusão de nível alto a partir dos modelos de classificação pela RF de cada técnica forneceu uma exatidão final de 90 a 100% no conjunto de teste.

Recentemente, outras técnicas espectroscópicas também são associadas à RF em problemas de classificação para discriminação quanto à origem. Mascellani e colaboradores ¹⁴⁰ classificaram 917 amostras de vinho por meio de RMN de 1H e modelos RF quanto ao tipo de vinho e variedades de uvas com exatidão de 93 a 99% para tipo de vinho e de 13 a 96% para as 13 variedades de uvas. RF também foi aplicada com sucesso para discriminação geográfica de amostras de azeite de oliva

extravirgem a partir de dados de espectroscopia NIR em sistema terahertz, com pulso de laser de Femtosegundo (>50 fs), de 1550 nm de comprimento de onda.⁹⁵ Para isso, os autores utilizaram algoritmo genético (GA) como método de seleção das variáveis de entrada dos modelos de classificação.

RF é naturalmente um método multiclasse.¹⁴ Mas há algumas propostas para utilização deste método como *ensemble* de classificadores de classe única (*one-class RF*).^{145–147} Este tipo de abordagem é comumente indicado para problemas de investigação de fraudes, principalmente em matrizes alimentícias.^{148–150} Neste caso, o modelo será treinado com as informações apenas da amostra autêntica, permitindo sua autenticação frente a diferentes tipos de adulteração.

Essa abordagem de classe única foi adotada em associação a dados de MIR e NIR por Santana, Borges-Neto e Poppi¹⁷ em problema de autenticação de matrizes alimentícias. *One-class RF* foi realizada a partir de espectros de HATR de amostras de óleo de prímula, e noz-moscada moída, quanto a adulterações. Um número otimizado pelos autores de 1000 *outliers* foram artificialmente gerados, a partir de um modelo PCA dos dados originais. Esses *outliers* são uniformemente distribuídos ao redor da classe-alvo, cobrindo a variação dos dados, para permitir uma simulação de tipos diferentes de adulterantes. Com os dados originais e *outliers* artificiais um modelo de classificação RF foi construído como um problema de duas classes, mantendo mesmo peso entre as classes. Os conjuntos de validação continham subconjunto de validação de amostras da classe-alvo e amostras com adulterações simuladas pelos autores com adulterantes comumente empregados e outros possíveis adulterantes. Os autores obtiveram alta sensibilidade e especificidade ($\sim 100\%$) com modelos *one-class* com RF e SIMCA, concluindo que RF teve um desempenho melhor devido ao número menor de amostras com erro de classificação.

Já para problemas quantitativos, ainda não há ampla aplicação de RF a partir de dados espectroscópicos. Isso pode estar associado à alta colinearidade desses dados e à tendência de superajuste comumente relatada em modelos não-lineares. RF não tem sido relatado como o método mais adequado em alguns estudos bem discutidos, como mostrou Dotto e colaboradores.⁷⁷ Neste artigo, os autores utilizaram nove métodos multivariados para previsão de carbono orgânico em 595 amostras de solo de diferentes regiões brasileiras a partir de dados de NIR no modo de reflectância. Os autores realizaram uma importante avaliação de diferentes métodos de

processamentos a partir dos desempenhos dos modelos. Entretanto, entre os métodos de regressão RF foi o que apresentou menor capacidade preditiva, sendo discutido ainda estudos anteriores com resultados semelhantes. Embora o foco principal do artigo sejam as técnicas de pré-processamento e seus efeitos nos resultados dos modelos preditivos, os autores destacaram algumas características importantes da RF. A exatidão dos modelos de RF manteve-se em ausência de algum dado ou presença de *outliers*. Além disso não fez previsões além do intervalo dos valores de resposta para os dados de treinamento. Destacaram como pontos negativos a falta de robustez do modelo, que alcançou bom desempenho para conjuntos de calibração, mas não eram mantidos ao conjunto de validação e falta de transparência ou interpretabilidade.

As distribuições de propriedades do solo podem ser assimétricas ao invés de uma distribuição normal. Nestes casos, RF pode apresentar melhores resultados quando há relação não-linear, como foi observado na previsão de salinidade em solo, por Wang e colaboradores.⁹⁴ Foram adquiridos quatrocentos espectros de espectroscopia na região do visível e infravermelho próximo (VIS-NIR), de comprimento de onda de 400 - 2400 nm de amostras de solo, de região próxima a um lago salino no noroeste da China. Com os espectros pré-processados pela derivada com ordens fracionárias o método RF apresentou melhores modelos ao ser comparado com modelo pelos mínimos quadrados parciais, obtendo-se a previsão de salinidade com coeficiente de determinação (R^2) de 0,90 a 0.97.

Também utilizando a técnica VIS-NIR, Santana, Souza e Poppi¹⁸ construíram modelos RF para previsão de parâmetros de qualidade a partir de um grande conjunto de dados de solo ($n=3854$). Os parâmetros matéria orgânica do solo (MOS), capacidade de troca de cátions (CTC), soma das bases (SB), teores de areia e argila foram modelados com RF a partir de espectros VIS-NIR (comprimento de onda de 400 - 2500 nm), com valores de erro médio estatisticamente melhores, comparados a um modelo pelo método PLS.

Há uma premissa comum a uma grande maioria dos trabalhos descritos acima, que é a interpretabilidade dos modelos de RF, muitas vezes compreendida como uma limitação à aplicação do método aos dados do ponto de vista químico. É comum atribuírem ao método RF o termo de caixa preta,¹⁵¹ ou caixa cinza, em analogia a sistemas de informação no qual um sistema de caixa preta é aquele em que somente

é possível visualizar as variáveis de entrada e de saída, sem conhecimento da parte interna de execução do algoritmo. Os sistemas de caixa cinza são aqueles em que é possível visualizar parcialmente como são feitos os cálculos pelo algoritmo. Também é comum a aplicação do RF por meio de pacotes com as funções internas, ou com sistemas de *toolbox* com interfaces, sendo aplicados geralmente com os parâmetros padrão adotados pelo desenvolvedor, sem maiores detalhes. Existem alguns trabalhos com maiores detalhamentos do método, aprimoramentos e adaptações,^{15,17,21,75,76,145–147,151–155} mas poucos são direcionados aos problemas químicos^{21,75,76,155} e aos dados espectroscópicos nos problemas do ponto de vista químico.^{15,17}

Neste sentido, um ano após a introdução do método, Breiman lançou um relatório técnico com descrições mais detalhadas, intitulado “*Looking inside the black box*”.¹⁵¹ Outro relatório técnico foi lançado três anos depois, por Chen, Liaw e Breiman¹⁵² com adaptações para utilizar RF em problemas de classificação em que há grande desbalanceamento entre as classes, uma alternativa denominada floresta aleatória balanceada. E em 2005, Livingston¹⁵⁴ reportou implementação de cálculo de variáveis importantes da RF no compilador de linguagem Java (Weka). Khoshgoftaar e colaboradores¹⁵⁶ também implementaram RF no compilador Weka, apresentando efeitos de diferentes números de variáveis como entrada *mtry* e números de árvores em modelos com classes desbalanceadas, além de avaliar robustez do método.

No contexto de complementos ou adaptações, Strobl e colaboradores¹⁵³ avaliaram o método de calcular importância das variáveis pelo índice de importância de Gini, proposto por Breiman.¹⁴ Este método é amplamente empregado para seleção de variáveis e interpretação do modelo RF. Entretanto, os autores consideram que há presença de viés, quando variáveis possíveis de serem preditoras em um modelo apresentem variabilidade em escala de medidas. Por meio de dados simulados os autores propuseram maneiras de minimizar este risco com uma técnica de amostragem sem reposição nas árvores de decisão.

Dessa forma, tem surgido alternativas para identificar e selecionar variáveis em modelos RF, como Granitto e colaboradores¹⁵⁷ que associaram a técnica de eliminação recursiva de variáveis aos modelos RF (RF-RFE, do inglês Random Forest-Recursive Feature Elimination) com dados de espectrometria de massa de reação de transferência de prótons (PTR-MS) de produtos agroindustriais. Lovatti e

colaboradores¹⁵ propuseram identificação de variáveis de acordo com a frequência de seleção pelo índice de Gini para interpretar as regiões de espectro de RMN de ^1H e ^{13}C em amostras de petróleo bruto. Em um trabalho posterior Lovatti e colaboradores¹⁶ realizaram seleção de variáveis e redução de dimensionalidade em dados de entrada em modelo RF. Por fim, Nembrini¹⁵⁸ apresentou um método de avaliação de medida de importância de variáveis por meio de floresta aleatória não supervisionada.

Uma outra adaptação proposta, mas de baixa discussão na química é a floresta aleatória não supervisionada. Apesar de ser naturalmente supervisionada, RF foi proposta como método não-supervisionado por Shi e colaboradores.¹⁵⁹ A partir disso tem sido sugerido como uma alternativa para reconhecimento de padrões em relações tanto lineares quanto não-lineares.^{21,160–164} O método RF não supervisionado utiliza a premissa de existir uma estrutura nos dados que permita a distinção entre dados originais e um conjunto de *outliers* artificiais. Essa abordagem tem sido empregada em estudos na área de saúde,¹⁶¹ geoquímicos,¹⁶⁰ meio-ambiente¹⁶⁴ e astronomia,^{162,163} mas ainda é pouco explorado em química analítica.²¹ É um método que pode ser uma importante ferramenta nos casos de uma triagem, visando a uma investigação inicial das semelhanças químicas entre amostras das quais não se tenha informações que as caracterizem.

RF não supervisionada pode ser uma importante alternativa para reconhecimento de padrões em alguns campos de estudo, como por exemplo na bioespectroscopia, em que muitas vezes a PCA não é suficiente para se conseguir uma resposta exploratória, frente à complexidade da informação. Além disso, muitos problemas neste campo de pesquisa tem o desafio de extrair informação de biofluidos para compreender as alterações bioquímicas provocadas pela doença ou por algum agente infeccioso. Mas geralmente passam por dificuldade de se conseguir um número satisfatório de amostras para investigação. Afanador e colaboradores²¹ demonstraram a potencialidade da RF não supervisionada para dados de origem biológica, comparando-a com PCA. Recentemente, esse método foi crucial para identificar estrutura em dados de espectroscopia MIR a partir de saliva em estudo de diagnóstico para COVID-19, em um trabalho que integra as áreas Ciências da Saúde e Ciência Exatas.²⁰ Já o método RF supervisionado, apesar de algumas aplicações com sucesso na bioespectroscopia, com técnicas espectroscópicas como RMN de ^1H ,^{138,139} e SERS-Raman,^{141–144} há poucas aplicações neste campo com a

espectroscopia infravermelho.^{123,165} Apesar disso, tem sido apontado como potencial ferramenta nesse campo de pesquisa.^{19,37}

Como pôde ser evidenciado, RF é um método versátil e eficiente, que vem sendo mais compreendido e adaptado com sucesso para diferentes abordagens em Química. Mas sua potencialidade em alguns campos de estudo, como a bioespectroscopia, ou outros campos que usam espectroscopia de absorção na região do infravermelho, encontram desafios que carecem de maior discussão na literatura.

2.4. Fundamentação Teórica

2.4.1 Floresta aleatória supervisionada

2.4.1.1 Árvore de decisão

A base da floresta aleatória é uma árvore de decisão. Uma árvore de decisão consiste em um conjunto de testes de decisão estruturados em árvore que segue uma abordagem “dividir-para-conquistar”. A ideia dessa abordagem de aprendizagem é dividir o espaço de dados em porções menores e mais susceptíveis de serem aprendidos por diferentes classificadores. Nesta técnica a população de amostra é dividida em dois ou mais conjuntos com partições recursivas com base em divisores significativos em suas variáveis de entrada. Um esquema de uma árvore de decisão é mostrado na Figura 2.8. O nó raiz representa a informação do conjunto de amostras ainda a ser dividido. Este nó é dividido em subnó ou nó de decisão. Esta divisão vai gerar subespaços, a partir do espaço original do conjunto de dados. O nó terminal, ou como é comumente chamado, nó folha, são os nós não divididos, devendo neste haver um conjunto mais puro possível das amostras para que se alcance uma resposta, que pode ser uma classe. Há ainda a terminologia para ramificação, que é quando há uma divisão criando uma subseção da árvore com um nó de decisão e nós terminais. Outra forma comum é chamar os nós de nó-pai ou ascendente, aquele nó que é dividido em outros subnós, e a estes chamar de nó-filho ou descendente.^{15,16,65}

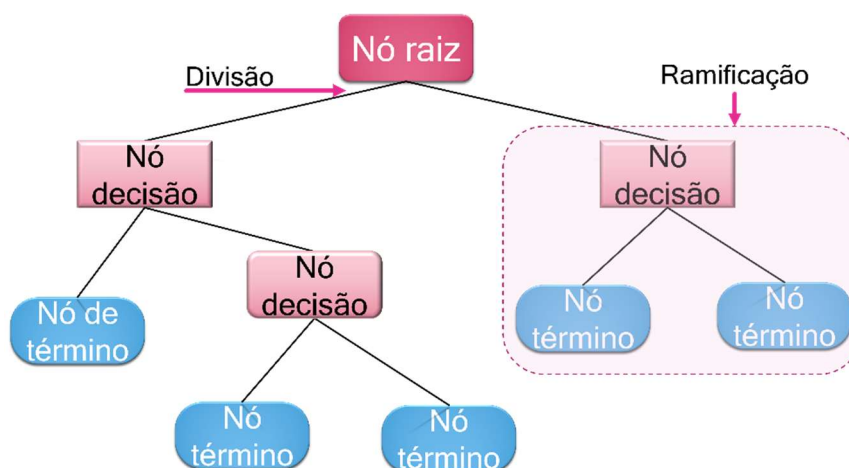


Figura 2.8 Um esquema geral de uma árvore de decisão.
Fonte: compilada pela autora.

Numa árvore de decisão para classificação, cada nó decisão é associado a um teste de uma variável, também chamado divisão, ou, seu termo em inglês, *split*. Os dados que “caem” neste nó serão divididos em subconjuntos de acordo com seus valores no teste da variável. Cada nó término (nó folha) está associado a um rótulo, que será atribuído às amostras que “caem” neste nó. Para predição, uma série de testes de variáveis é realizada a partir do nó raiz e o resultado é obtido quando se atinge um nó folha, como no exemplo dado na Figura 2.9, cujos valores são de um exemplo hipotético. O processo de classificação começa testando e dividindo os dados pela variável preditora x_{17} . As amostras pertencentes a diferentes classes com valor nesta variável menor do que o valor dado no teste, são direcionadas ao lado esquerdo, compondo um nó decisão, enquanto as com valor maior ou igual são direcionadas para o lado direito, formando um nó folha associado à classe 2. Novo teste é realizado em outra variável, no caso x_3 , e assim sucessivamente até formar os nós terminais (folhas).

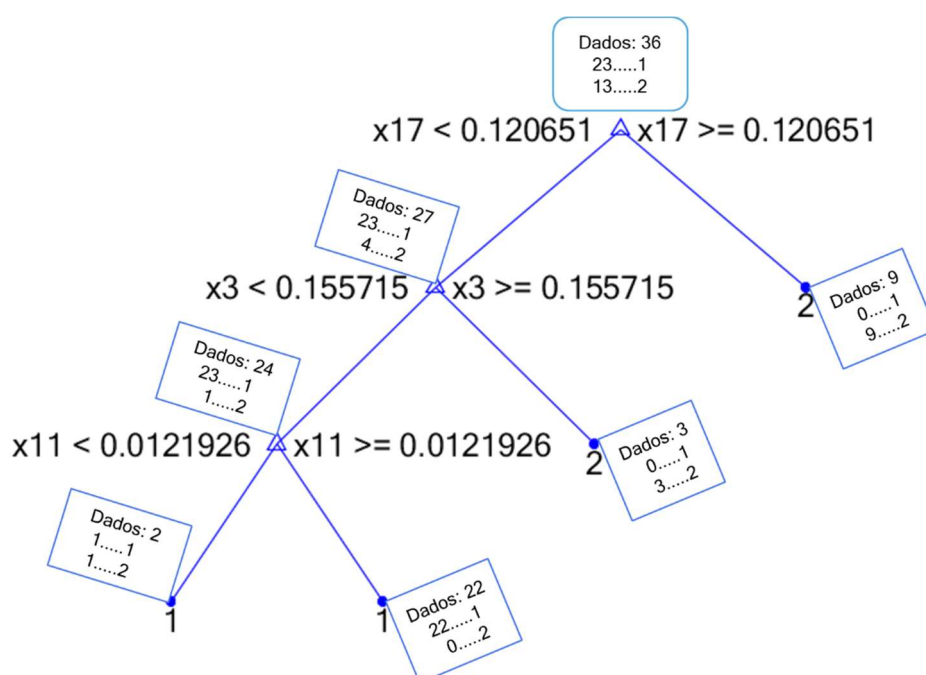


Figura 2.9 Exemplo esquemático de divisão de uma árvore de decisão.
Fonte: compilado pela autora.

Os algoritmos de decisão utilizam uma métrica para selecionar as variáveis que participam destas divisões. Os algoritmos iniciais de árvore de decisão eram habilitados para analisar apenas variáveis categóricas e selecionar com base em uma taxa de ganho de informação calculada a partir da entropia, considerando que ao se dividir um conjunto de treinamento em subconjuntos a entropia pode ser reduzida. O valor dessa redução é o ganho de informação. É selecionada a variável com a maior taxa de ganho, entre as variáveis com ganhos de informação acima da média. Outros métodos posteriores foram habilitados a utilizar dados numéricos.³⁵

Já no algoritmo CART (*Classification and Regression Trees*)⁶⁹ que é a base da floresta aleatória, as divisões podem ser feitas com base em dados categóricos ou dados numéricos, como no exemplo dado na Figura 2.9, sendo método padrão de seleção o índice de Gini.¹⁴

Vale ressaltar que as árvores de decisão estão relacionadas com o problema de sobreajuste, como mencionado anteriormente. Na proposta da RF é formado um *ensemble* com arvores de decisão sem ocorrer a poda (*unpruned tree*), para manter o viés baixo. Enquanto a exatidão é garantida pela baixa correlação entre os classificadores, dada pela aleatoriedade.^{14,35}

2.4.1.2 Ensemble

Um *ensemble*, que também pode ser chamado de aprendizagem em conjunto, ou sistema de classificadores múltiplos, contém um número de classificadores chamado de classificador base, que são usualmente gerados a partir do conjunto de treinamento com uso de um algoritmo base, podendo ser uma árvore de decisão, uma rede neural *etc.* Esse modelo de base pode ser de um único tipo, o que produz um *ensemble* homogêneo. Em outro sentido, se são combinados modelos de base de diferentes tipos/algoritmos forma-se um *ensemble* heterogêneo. Nesse caso, como a base não pertence a um mesmo tipo, não há um consenso em denominar o modelo individual como modelo base. Por isso pode ser encontrado sendo mencionado como “modelo individual” ou “componente individual”.^{35,166}

Geralmente, um *ensemble* possui maior habilidade de generalização do que um modelo de base. Também é comum descrever os modelos de base como “modelos fracos”, que seriam pouco melhores do que um resultado aleatório, e o *ensemble* como o método capaz de impulsionar esses modelos mais fracos para classificadores mais fortes que podem gerar previsões com maior exatidão. Uma estrutura básica de um ensemble é representada na Figura 2.10. A ideia é construir muitos classificadores e então combinar suas saídas de modo que o desempenho medido na resposta final (y) seja melhor do que o desempenho de cada classificador.^{14,35,69}

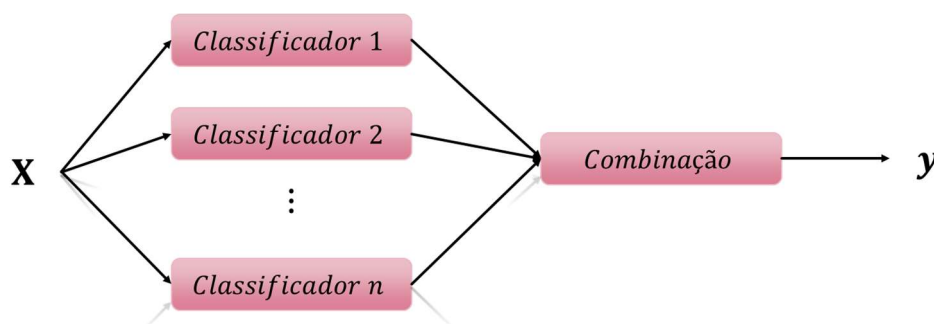


Figura 2.10 Estrutura básica de métodos ensemble.
Fonte: Elaborado pela autora, adaptado de Zhou, Z.³⁵.

Dentro dos três segmentos de discussão dos métodos ensemble, ou seja, combinação de classificadores, conjuntos de classificadores fracos e mistura de especialistas, o primeiro é o mais estudado no campo de reconhecimento de padrões. Os pesquisadores nessa área geralmente trabalham com classificadores fortes e empregam regras de combinação para obter classificadores combinados mais fortes.

Já o segmento conjunto de classificadores fracos é abordado principalmente na área de aprendizagem de máquina. Nesse segmento, os trabalhos geralmente abordam classificadores fracos (rede neural, árvore de decisão etc.) projetados em algoritmos iterativos para impulsionar o desempenho de fraco para forte. Como regra de combinação é mais comum empregar o voto majoritário em conjunto de modelos de classificação. Cada modelo vota em uma classe e a classe com mais votos é atribuída como sendo o rótulo da amostra de teste.^{35,166}

A aplicação de *ensembles* é útil em conjuntos com grande volume de dados, aos quais pode-se usar uma estratégia de dividir em subconjuntos e construir classificadores com diferentes subconjuntos. Ao final esses modelos são combinados por meio de uma regra de combinação. Já no caso de haver pequeno volume de dados, pode-se empregar técnicas de reamostragem na criação de subconjuntos aleatórios em relação aos dados disponíveis, utilizando cada subconjunto para treinar diferentes classificadores. A combinação desses classificadores forma um *ensemble* com desempenho melhor.

Um *ensemble* também pode ter uma estrutura conhecida como análise de população de modelos (MPA, do inglês, *Model population analysis*), proposto inicialmente como método de seleção de variáveis.¹⁶⁷ MPA é uma estrutura que aplica técnica de amostragem aleatória pelo Monte Carlo e técnicas estatísticas para comparação da população de submodelos.^{168,169} É caracterizado em três etapas principais (Figura 2.11): (1) geração aleatória de N subconjuntos de dados pela amostragem com Monte Carlo; (2) construção de um submodelo em cada subconjunto de dados; (3) análise estatística de algum parâmetro de interesse do submodelo. Pela amostragem Monte Carlo podemos adquirir N subconjuntos de dados extraídos de N amostragens a partir de um conjunto original $(\mathbf{X}_{(m,p)}; \mathbf{y}_{(m,1)})$. Para cada subconjunto de dados $(\mathbf{X}_{sub(m',p)}; \mathbf{y}_{sub(m',1)})$, um submodelo pode ser construído usando um método selecionado, que pode ser de classificação ou de regressão. Então, todos esses submodelos podem ser combinados. Ao final, todos os N submodelos são mutuamente diferentes, mas com o mesmo objetivo de prever uma resposta para $\mathbf{y}$.^{167–169}

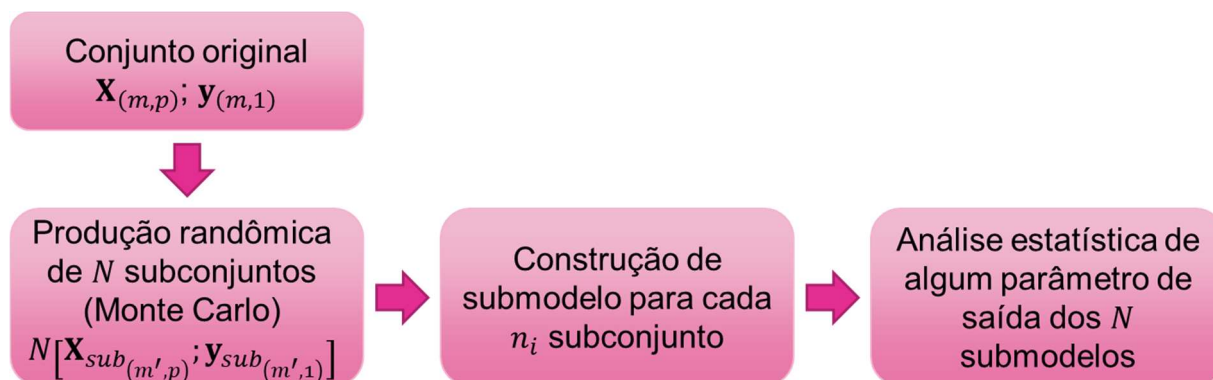


Figura 2.11 Esquema simplificado de etapas de MPA para classificação. Fonte: adaptado de Li et al.¹⁶⁹

Os submodelos podem ser construídos com a aleatoriedade empregada no espaço das amostras, espaço das variáveis, espaço paramétrico ou espaço dos modelos. É importante destacar que pelo MPA a resposta de interesse não é só um valor singular, mas uma distribuição dessa resposta.¹⁶⁹ Para gerar a população de submodelos, a amostragem aleatória (Monte Carlo) é geralmente aplicada por diferentes técnicas para se obter diferentes subconjuntos de amostras, ou até mesmo subconjuntos de variáveis, sendo mais empregado em MPA as técnicas *Jackknife* (JNS), que faz amostragem sem reposição de amostras, *bootstrap*, que faz amostragem com reposição, e amostragem por matriz binária (BMS, do inglês, *binary matrix sampling*), que utiliza o sistema binário 0 ou 1 para marcar espaços como não selecionados ou selecionados, respectivamente. Outra forma de amostragem empregada em MPA é a amostragem por permutação, que difere de uma amostragem aleatória porque na permutação é destruída a conexão entre as amostras e resposta de interesse. Em MPA essa amostragem é mais aplicada para obter uma variedade de modelos “errados” e se obter uma avaliação por exemplo da importância de variáveis, comparando-se as distribuições de algum parâmetro dos modelos sem a permutação *versus* com a permutação, entre outros objetivos. A análise estatística com base na distribuição de um parâmetro de interesse é um ponto característico do método MPA. Geralmente, essa distribuição é descrita pelo valor médio, desvio padrão, mediana e intervalo de confiança. Os testes estatísticos empregados são testes comparativos, como teste t, quando este parâmetro tem distribuição normal, ou teste U de Mann-Whitney, para distribuição não-normal. Uma apresentação mais detalhada da estrutura do MPA está demonstrada na Figura 2.12.

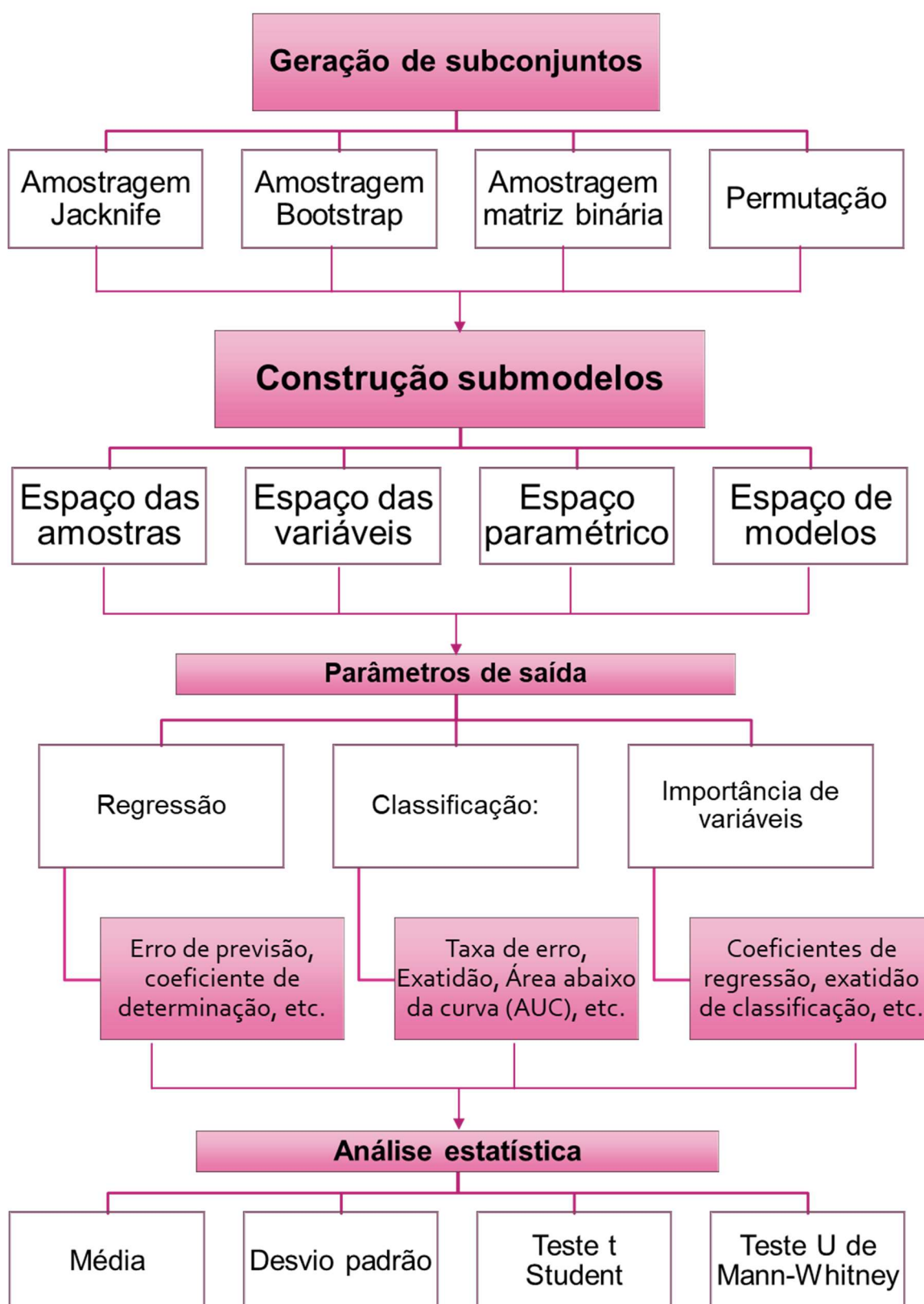


Figura 2.12 Fluxograma representando estruturas possíveis de MPA. Fonte: adaptado de Deng et al.¹⁶⁸

2.4.1.3 Floresta aleatória (RF)

Uma floresta aleatória é uma coleção de árvores preditoras (CART), construídas com vetores aleatórios identicamente distribuídos e independentes (Bagging). Cada árvore lança um voto unitário para a classe mais popular em dados de entrada (regra de combinação). Apesar de haver pontos em comum, como a aleatoriedade na amostragem e a combinação de modelos, a floresta aleatória difere de um MPA pela construção da RF ter como base um modelo de um tipo apenas (*ensemble* homogêneo) e a utilização do voto majoritário como regra de combinação.

Um modelo supervisionado de classificação, pela RF parte de um conjunto de dados relacionados à amostra, que podem ser dados espectroscópicos, ou outro tipo de informação, sendo a matriz de dados ($\mathbf{X}_{(n,p)}$) e um vetor de rótulo da classe ($\mathbf{y}_{(n,1)}$) contendo a informação química qualitativa. Subconjuntos de treinamento ($\mathbf{X}_{\text{trein.}(n',p)}$) e de teste ($\mathbf{X}_{\text{teste}(n'',p)}$) são adquiridos a partir dos dados originais pelas diversas técnicas de amostragem disponíveis (Figura 2.13, A). A partir do conjunto de treinamento são adquiridos os subconjuntos aleatórios *In-Bag* ($\mathbf{X}_{\text{In-Bag}(n',p')}$) e *Out-of-Bag* ($\mathbf{X}_{\text{OOB}(n_{\text{OOB}},p')}$) pelo algoritmo Bagging (Figura 2.13, B). O conjunto *In-Bag* contém $\sim 2/3$ das amostras, mas mesma dimensão de linhas do conjunto de treinamento. Consequentemente, OOB contém $1/3$ das amostras do conjunto de treinamento. Nesta etapa também ocorre a seleção aleatória de variáveis (p') pelo valor adotado para o parâmetro *mtry* (padrão = $\sqrt{\text{variáveis}}$). Para o crescimento da árvore ocorre ainda a seleção das variáveis preditoras entre p' a partir de um processo de validação cruzada (padrão = k-fold) com o conjunto In-Bag. Com esses passos, a primeira árvore é crescida sem ser podada (CART), isto é, seleção das variáveis preditoras, validação cruzada e treinamento com o conjunto In-Bag, seguidos de validação pela taxa de erro de OOB.

Todo o processo, começando com uma nova amostra *bootstrap*, ocorre iterativamente até o total de *nTrees* (Figura 2.13, B). O voto majoritário é utilizado como regra de combinação das respostas de cada classificador independente, para formar uma resposta final para cada amostra de treinamento. Todas as árvores preditoras da RF então construída são aplicadas para predição de classe de amostras no conjunto de teste (Figura 2.13, C) e avaliação do modelo.

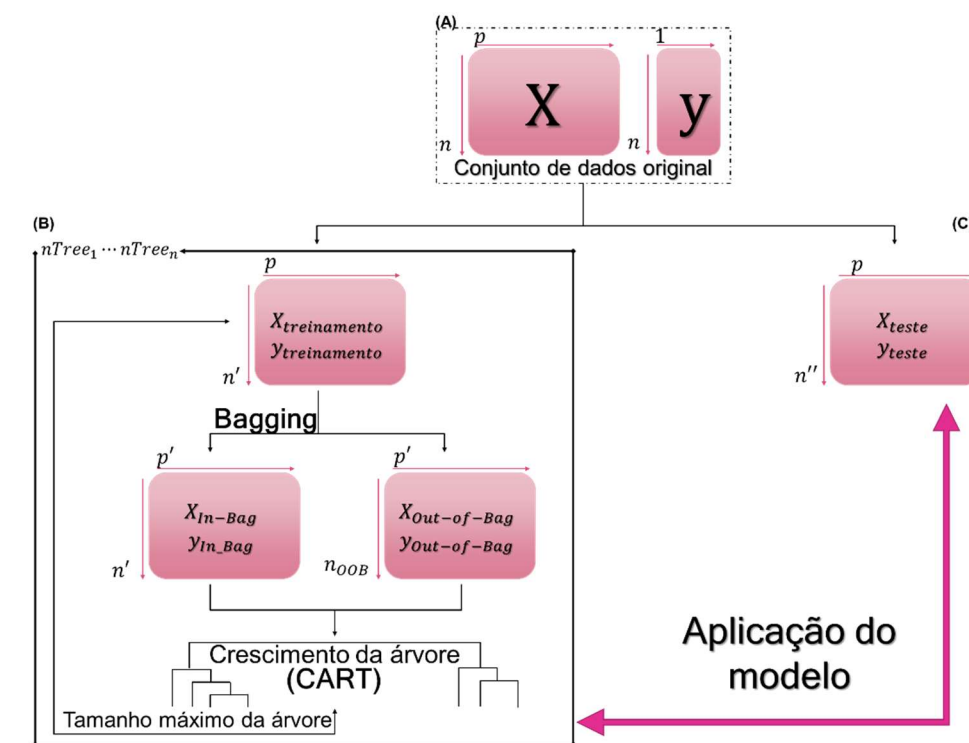


Figura 2.13 Fluxograma de crescimento de árvore de classificação ou de regressão em modelo de RF.

Fonte: compilado pela autora, adaptado de Breiman, 2011¹⁴.

Durante o crescimento da árvore, cada variável do subconjunto amostrado aleatoriamente (p') é analisada como possível preditora (nó) em um procedimento denominado pesquisa exaustiva. A partir do nó raiz, onde o subconjunto de variáveis é testado em busca de uma melhor divisão, seguindo-se para próximas variáveis com mesmo procedimento para formar os nós descendentes. Os algoritmos utilizam alguns critérios para minimizar a impureza do nó, e, conseqüentemente, maximizar a homogeneidade. Em um problema de classificação podem ser usados:^{14,166}

- erro de classificação incorreta (do inglês, *missclassification error*), uma medida da proporção de observações no nó que não são membros da classe majoritária naquele nó. Esta medida não é muito comumente empregada;
- índice de entropia, que também é chamado de entropia cruzada ou *deviance*, que é uma medida de desvio da impureza. Este índice é mais sensível do que *missclassification error* a mudanças na probabilidade do nó;
- regra de Twoing, um critério projetado para problemas multiclases, que prioriza a separação entre as classes com a divisão tratada como um problema binário, mas favorecendo divisões que mantêm classes relacionadas juntas;

- Índice de Gini, que é o critério mais comumente empregado, sendo o método padrão em vários algoritmos ou funções, incluindo CART. Tem a função de minimizar a impureza do nó e é mais sensível do que *missclassification error* a mudanças na probabilidade do nó.

O índice de impureza Gini pode ser definido pela Equação 01.

$$Gini(nó) = 1 - \sum_{i=1}^c \left[p\left(\frac{i}{nó}\right) \right]^2 \quad \text{Eq. 01}$$

Onde, c é o número de classes e $p\left(\frac{i}{nó}\right)$ é a fração dos registros pertencentes à classe i no nó.^{14,15} A decisão pela variável preditora é realizada pelo ganho de informação ao calcular a diferença entre o índice de Gini(nó) do nó ascendente (nó pai) e nó descendente (filho) (Equação 02).

$$Ganho_{Gini} = Gini(ascendente) - \sum_{nd=1}^n \left[\left(\frac{nd}{na}\right) * Gini(nd) \right] \quad \text{Eq. 02}$$

Onde, n é o número de nós descendentes, nd é o número total de eventos no nó descendentes e na é o número total de eventos do nó ascendente. Um exemplo desse cálculo de Gini para a árvore de classificação demonstrada na Figura 2.9 é apresentado no Anexo A.

Uma característica importante da RF é que à medida que mais árvores são adicionadas há uma resistência ao sobreajuste, produzindo um valor limite do erro de generalização. Isso é definido com base em teoremas matemáticos sobre a força dos *ensembles* e a estrutura das árvores. Ademais, o erro de generalização é usado para calcular força e correlação dos classificadores individuais, parâmetros usados por Breiman¹⁴ para expressar a exatidão do método. De acordo com sua descrição, a aleatoriedade inserida gera um modelo robusto para *outliers* e ruído, porque minimiza a correlação entre os classificadores enquanto mantém a força.

Outro ponto que deve ser destacado é que o erro de generalização é a taxa de erro do classificador OOB no conjunto de treinamento. As características das amostragens dados pelo Bagging fazem com que a estimativa de OOB não seja viesada e seja equivalente a uma validação por um conjunto de teste de mesmo tamanho do conjunto de treinamento. Dessa forma, com a estimativa de erro OOB é satisfeita a validação equivalente a um teste externo, dispensando um conjunto separado para este fim.¹⁴

2.4.2 Floresta aleatória de classe única (OCRf)

Como um método principalmente de classificação binária, os modelos RF devem ser treinados com amostras de pelo menos duas diferentes classes. Já modelos de classe única são treinados apenas pela classe-alvo.^{148–150} Portanto, RF como modelo de classe única inicia-se pela geração de dados sintéticos como possíveis *outliers*, ou seja, possíveis variações de dados não-pertencentes à classe-alvo. A partir da concatenação de um conjunto de dados artificiais e seu conjunto original é que é treinada a RF. Para o conjunto de teste são adicionadas as amostras da classe-alvo, separadas para o conjunto de teste, e amostras não-pertencentes à classe-alvo, ou suspeitas de fraude, de seu conjunto original.

Um fluxograma com a representação esquemática da construção de um modelo de classe única pelo RF é demonstrado na

Figura 2.14. Nele, a partir de um conjunto de treinamento contendo somente amostras da classe alvo são gerados *outliers* artificiais. Uma versão sintética criada a partir da distribuição dos dados originais, pela decomposição de uma PCA por valores singulares (SVD) e um valor de desvio-padrão (geralmente 3). Para geração dessa versão sintética, o conjunto de treinamento contendo apenas a classe alvo é centrado na média para transformação da base vetorial pela PCA. Em seguida, n *outliers* sintéticos são geradas randomicamente na mesma dimensão das componentes principais, ou seja, na matriz de scores (**T**). Ainda na dimensão das PCs, os semieixos de uma elipsoide são calculados para compor os limiares de distribuição dos dados originais, adotando-se um limite de confiança de 3 desvios padrão. Os *outliers* artificiais são distribuídos acima deste limite e transformados para a base vetorial dos espectros por meio da soma de um vetor de média inicial ao produto entre as matrizes de dados sintéticos randômicos e de *loadings*. O valor de n para quantidade de amostras no conjunto de *outliers* artificiais, tem sido adotado como o mesmo tamanho do conjunto original ou um valor que expressa uma quantidade significativamente maior do que este conjunto.¹⁷ Por fim, os dados de treinamento e sintéticos são concatenados e um vetor de rótulo de classes é criado contendo as classes “original” e “sintético” para os respectivos conjuntos. A seguir, é realizado todo o processo iterativo de construção de uma RF de classificação binária a partir deste novo conjunto concatenado. No conjunto de teste externo há amostras da classe alvo e amostras não pertencentes à classe alvo, por exemplo amostras adulteradas. A este conjunto

de teste são previstas as classes pela RF construída. O modelo é avaliado pelos parâmetros de OOB e de teste.

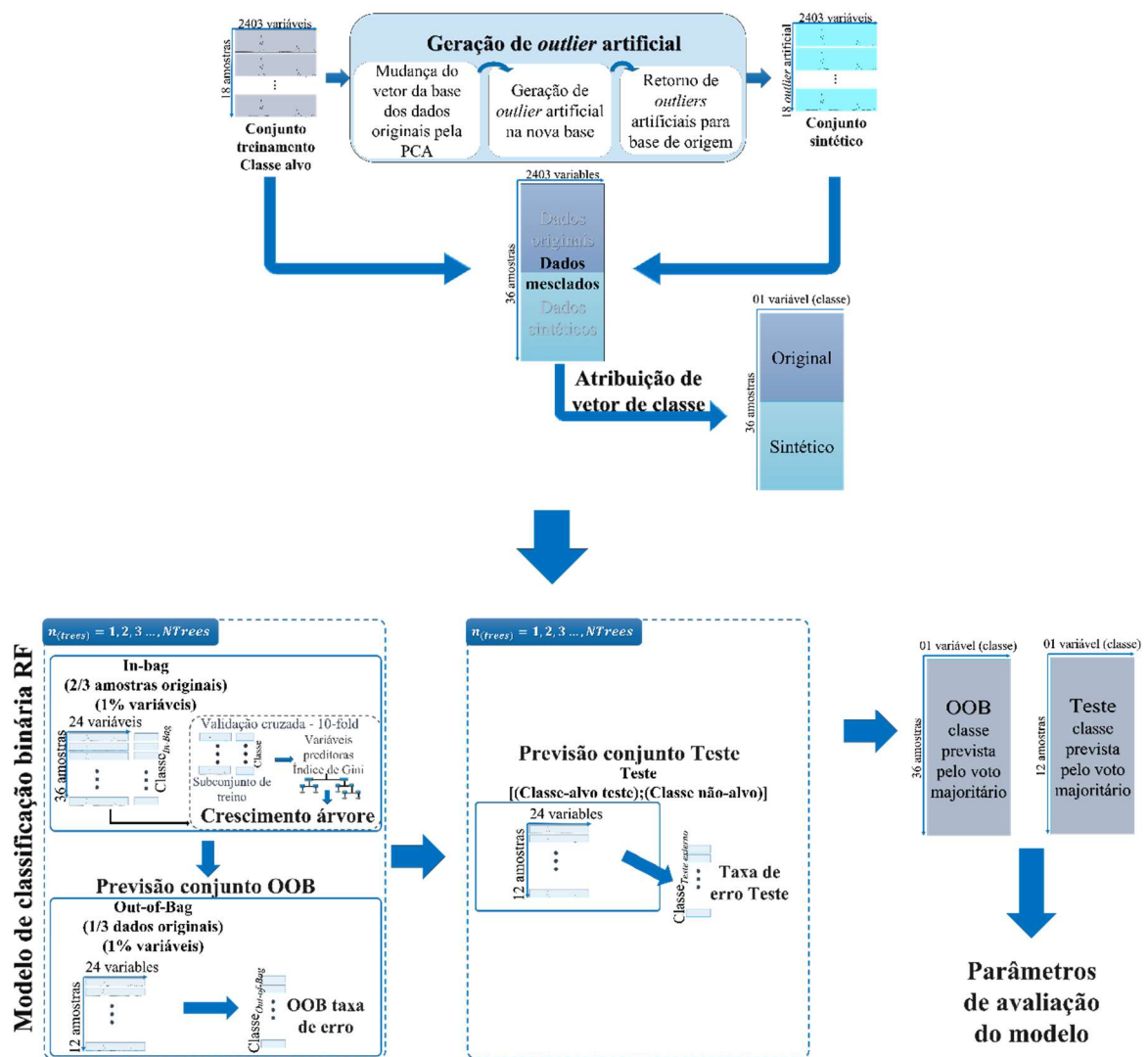


Figura 2.14 Fluxograma com passos para construção de modelo de floresta aleatória de classe única (OCRF). Fonte: compilado pela autora.

2.4.3 Floresta aleatória não supervisionada (URF)

O método URF utiliza um modelo RF de classificação binária para distinguir um conjunto original de amostras não rotuladas, ou seja, que não apresentam uma informação de resposta, de uma versão aleatória deste mesmo conjunto, sinteticamente criada. O método URF descrito neste trabalho é uma adaptação, de acordo com metodologia de Shi e Horvatgh¹⁵⁹ e Afanador e colaboradores.²¹ Este método usa o princípio de que havendo alguma estrutura nos dados, estes serão distinguíveis de uma versão sintética e aleatória de si mesmos.^{21,159} Como outros métodos de classificação, a taxa de erro OOB é a métrica para validar a RF como viável para identificar uma estrutura de conjunto de dados. Altos níveis de taxa de erro ($\approx 50\%$) implicam ausência de estrutura no conjunto de dados original.¹⁷⁰

A partir deste modelo URF, semelhanças e diferenças de pares das amostras não rotuladas podem ser avaliadas por meio de matriz de dissimilaridade e análise multidimensional. Primeiro, um conjunto de dados sintético é gerado aleatoriamente com base na distribuição dos dados originais não rotulados. A partir de um modelo RF de classificação binária com as matrizes de dados original e sintético concatenados é possível identificar se há ou não há uma estrutura nos dados. Havendo uma estrutura que permita discriminar os dados originais de uma versão sintética aleatória, uma matriz de proximidade é usada para identificar os pares de amostras que seguiram as mesmas rotas e concluíram em mesmo nó terminal, a partir de árvores de decisão independentes. Além disso, esta análise pelos pares de amostras pode ser facilmente interpretada pela projeção das amostras em dimensão reduzida, via análise de coordenadas principais (PcoA), gerada a partir de uma matriz de dissimilaridade.

A análise de coordenadas principais é um tipo de dimensionamento multidimensional métrico (MDS, do inglês, *multidimensional scaling*) que avalia os pares de amostras com distâncias numéricas. PCoA representa as distâncias entre as amostras, atribuindo uma coordenada a cada amostra em um espaço de baixa dimensão. O primeiro passo é a construção de uma matriz de dissimilaridade. Embora seja baseado em uma matriz de dissimilaridade, quando se usa a distância Euclidiana a interpretação da PcoA pode ser semelhante à de uma PCA. Isso ocorre porque a PcoA reduz a dimensão de uma matriz de distância, baseado em uma equação de autovalor. Primeiro calcula uma matriz centralizada. A matriz centralizada é então decomposta em seus autovalores e autovetores por uma decomposição de valores

singulares. Os autovetores são padronizados dividindo-os pela raiz quadrada dos respectivos autovalores. Os autovetores padronizados são produzidos como eixos de coordenadas principais (Pco) de um gráfico bidimensional. As distâncias entre os pontos deste gráfico aproximam-se das distâncias na matriz de dissimilaridade. Essa análise também é chamada de análise por ordenação de dados multivariados e pode ser utilizada com base em qualquer função de distância. Ambos, PCA e PcoA usam redução de dimensionalidade, mas enquanto PCA converte correlações em um gráfico bidimensional a partir da máxima variância pelas PCs, PcoA extrai dimensões que representam máximas distâncias entre os pares de amostras.^{21,171}

Todas as etapas de construção de modelo URF são mostradas na

Figura 2.15. Uma versão do código, contendo algoritmo e exemplo para construção de modelo URF, em MATLAB, está disponível na plataforma GitHub (<https://github.com/MarciaHelenaCassago/URF-model>). Dado um conjunto hipotético de amostras não rotuladas ($\mathbf{X}_{(48,1676)}$), de forma resumida, os procedimentos para construção de um modelo URF são:

- geração de um conjunto de *outliers* artificiais ($\mathbf{X}_{art(48,1676)}$), com base na distribuição dos dados originais não rotulados, como descrito no item 2.4.2. Este conjunto artificial e aleatório pode ser construído dessa forma, ou pela permutação das variáveis do conjunto original;
- concatenação das matrizes conjunto de dados original e sintético, formando o conjunto de dados de entrada para o modelo RF ($\mathbf{X}'_{(96,1676)}$), ;
- criação de vetor contendo duas classes: "sintético" e "original" para as respectivas amostras ($\mathbf{y}_{(96,1)}$),;
- construção do modelo RF a partir deste conjunto de dados completo, como um problema supervisionado de classificação binária;
- identificação das amostras alocadas nos nós folhas das árvores no modelo RF;
- construção de matriz de proximidade;
- redução de dimensão pelo modelo PCoA;
- interpretação gráfica do modelo PCoA.

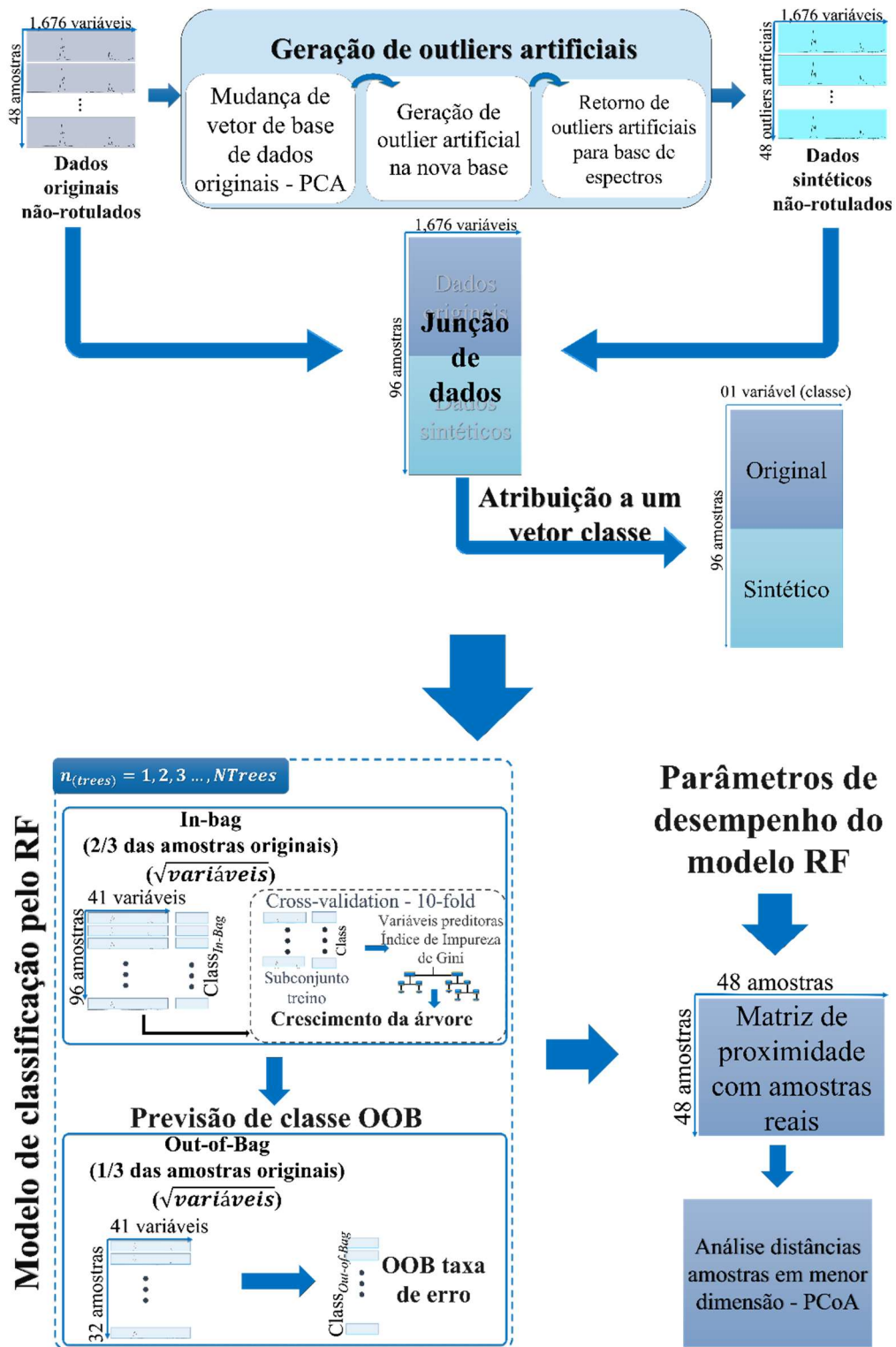


Figura 2.15 Fluxograma para construção de modelo floresta aleatória não-supervisionado (URF).
Fonte: compilado pela autora.

Os métodos não supervisionados não requerem métricas de qualidade para avaliar o desempenho de um modelo. No entanto a RF não supervisionada é dependente de uma avaliação dos erros no modelo RF supervisionado realizado em umas das etapas desse método. Se este modelo apresentar um índice de erro $\geq 50\%$, sugere-se que não há estrutura nos dados capaz de distinguir uma versão original de sua versão sintética e aleatória.²¹ Dessa forma, assim como em todo modelo supervisionado, a RF é avaliada pelo seu conjunto de testes. No caso da URF essa avaliação é realizada para o conjunto de testes OOB e no caso da RF ou OCRF também para o conjunto de teste externo.

2.4.4 Parâmetros de avaliação dos modelos

Os desempenhos dos modelos de classificação são avaliados por meio dos conjuntos de testes externos e diversas métricas. Essas métricas são determinadas a partir de uma organização dos resultados em uma tabela de contingência, ou matriz de confusão. Nesta tabela os números de amostras classificadas em cada conjunto são dispostos entre as classes reais, que são representadas nas colunas, e as classes estimadas pelo modelo, representadas nas linhas. Uma tabela de contingência para um problema de classificação de duas classes é apresentada na Tabela 2.3, assim como a terminologia dada aos valores que compõem a tabela, como Verdadeiro Positivo (VP), Verdadeiro Negativo (VN), Falso Positivo (FP) e Falso Negativo (FN).^{11,172}

Tabela 2.3. Demonstração de Tabela de contingência em problema de duas classes.

		Classe Real		
		1	2	
Classe prevista	1	VP	FP	VP+FP
	2	FN	VN	FN+VN
		VP+FN	FP+VN	

Antes de definir cada um dos termos acima, vamos considerar uma metodologia necessária para nortear a tabela de contingência, o teste de hipóteses. O teste de hipóteses estatísticas é um método que auxilia na tomada de decisões a respeito de uma ou mais populações, com base nas suposições (hipóteses estatísticas) e parâmetros da distribuição de uma amostra dessa população. Essas

suposições são testadas para serem confirmadas ou rejeitadas. A hipótese nula (H_0) é atribuída à hipótese a ser testada. A hipótese alternativa (H_a) é a hipótese aceita quando há indícios estatísticos suficientes para que seja rejeitada a hipótese nula. Junto ao teste de hipóteses é analisada a probabilidade de se tomar a decisão incorreta. Ou seja, ao se rejeitar H_0 sendo ela verdadeira, ocorre o erro do Tipo I (Falso Positivo). Da mesma forma, ao não se rejeitar H_0 sendo ela falsa, ocorre o erro do tipo II (Falso Negativo). O nível de significância do teste (α), ou nível de confiança, é a probabilidade máxima de ocorrência do erro do Tipo I. Isso significa que com o nível de significância geralmente adotado de $\alpha = 0,05$, a probabilidade de se rejeitar incorretamente H_0 é de 5%. Em contrapartida, a probabilidade de se tomar a decisão correta é dada por $(1 - \alpha) \cdot 100\%$. Já a probabilidade de ocorrência de erro do tipo II é demonstrada pelo símbolo β e é conhecido como poder do teste estatístico.¹⁷²

Um erro do tipo I (α) é a proporção de amostras pertencentes à distribuição nula H_0 que são incorretamente previstas como membro da distribuição alternativa H_a . Enquanto o erro do tipo II (β) é a proporção de amostras pertencentes à distribuição alternativa H_a e são incorretamente previstas como membros da distribuição nula H_0 .

Para um caso hipotético de classificação de 190 amostras de um biofluido para diagnóstico de uma determinada doença, consideramos dois grupos de amostras entre indivíduos doentes (Positivo) e saudáveis (Negativo), tem-se um problema de 2 classes, representado na tabela de contingência (Tabela 2.2), a hipótese nula e alternativa são descritas como: H_0 = a amostra não apresenta características reconhecidas da doença e pertence à classe 2, Negativo. H_a = a amostra apresenta as características reconhecidas da doença e pertence à classe 1, Positivo.

Pela Tabela 2.2, as decisões a respeito dessas hipóteses estão relacionadas da seguinte forma: α = probabilidade de um erro tipo 1 se uma amostra fizer parte da classe nula = $FP/(VN + FP)$. β = probabilidade de um erro tipo 2 se uma amostra fizer parte da classe alternativa = $FN/(VP + FN)$

Portanto, neste caso hipotético, havendo 90 amostras de indivíduos doentes e destas 86 sejam corretamente classificadas, e 100 amostras saudáveis, ou seja,

da classe negativa, das quais foi previsto pelo modelo 88 corretamente, haveria 86 VP, 4 FN, 88 VN e 12 FP, conforme demonstrado na Tabela 2.4.

Tabela 2.4 Demonstração de Tabela de contingência em problema hipotético de classificação de 90 amostras de biofluido positivo para determinada doença e 100 amostras negativas.

		Classe Real		
		Positivo	Negativo	
Classe prevista	Positivo	86	12	98
	Negativo	4	88	92
		90	100	

Dessa forma, podemos calcular os erros tipo I (α) e erro do tipo II (β) para as decisões acima:

$$\alpha = FP/(VN + FP) = 12/(88 + 12) = 12/100 = 0,12$$

$$\beta = FN/(VP + FN) = 4/(86 + 4) = 4/90 = 0,044$$

Dependendo do problema sendo investigado pode ser adequado minimizar o erro do tipo I, mesmo em detrimento do erro do tipo II, e vice-versa. Por exemplo, em uma análise forense de adulteração de diesel pode ser mais interessante minimizar o erro do tipo I, evitando-se assim prejuízos de uma possível reprovação de um combustível corretamente comercializado.

Para Brereton¹⁷³ é mais prudente minimizar um erro do tipo II em um estudo de diagnóstico na área de medicina. Nesse caso, um erro do tipo I, de uma hipótese nula de paciente saudável sendo erroneamente rejeitada, pode provocar que um paciente saudável tenha que realizar exames para confirmar uma enfermidade. No sentido contrário, o erro do tipo II, em que a hipótese nula de que o paciente seja saudável não seja rejeitada, sendo ela falsa, pode acarretar maiores prejuízos visto que um paciente realmente doente não irá prosseguir com exames e tratamentos necessários.

Compreendidas e definidas as hipóteses, a partir da matriz de confusão são calculados os parâmetros de desempenho do modelo de classificação. As principais métricas para essa avaliação são a exatidão (EXAT), sensibilidade (SENS), especificidade (ESPEC), ou seletividade, taxa de falso positivo (TFP) e precisão (PREC). A sensibilidade descreve a habilidade do modelo de classificar

corretamente as amostras pertencentes à classe sendo avaliada, ou seja, é a taxa de verdadeiros positivos. Especificidade caracteriza a habilidade de o modelo rejeitar corretamente as amostras não pertencentes à classe sendo avaliada, isto é, taxa de verdadeiros negativos. A precisão é uma métrica que descreve a probabilidade de amostras classificadas pelo modelo como pertencentes a uma classe, efetivamente serem pertencentes a esta classe. Enquanto exatidão é a razão das amostras corretamente classificadas entre todas as amostras da classe. A exatidão da classificação em um modelo binário é a média da sensibilidade e especificidade.¹⁷²⁻¹⁷⁴

As equações para calcular os parâmetros de desempenho descritos acima, comumente apresentados na literatura para validação de modelos de classificação, estão descritos na Tabela 2.3. Entre eles, o valor de F1-score é uma métrica utilizada para representar estatisticamente o conjunto de parâmetros de desempenho por meio de um único valor escalar. Esse valor escalar é uma média harmônica da precisão e sensibilidade, sendo 1 o melhor escore, consequentemente um excelente modelo e 0 o pior escore.¹⁷²⁻¹⁷⁴

Outra métrica que tem sido apresentada na literatura para demonstrar a qualidade dos modelos é o coeficiente de correlação de Matthews (MCC). Este parâmetro é adquirido a partir da matriz de confusão e tem sido utilizado como parâmetro de qualidade, principalmente em problemas com desbalanceamento entre as classes. Nestes casos, não é prudente utilizar a exatidão como principal métrica de qualidade do modelo. A partir da tabela de contingência MCC pode ser compreendido como a razão da diferença entre o produto da diagonal e os termos fora dela pela raiz quadrada do produto das somas das linhas e colunas. Vale destacar que essa forma de determinar este parâmetro é aplicado a um problema com duas classes. O valor de MCC varia de -1 a +1. Um valor de MCC próximo a -1 indica total discordância entre classe prevista e classe real. MCC próximo a zero indica que a predição não é melhor do que uma predição aleatória para seus respectivos dados. Em contrapartida, um valor próximo a +1 indica excelente desempenho e confiabilidade do modelo.^{172,175,176}

Tabela 2.5 Principais parâmetros de desempenho para validação de modelos de classificação.

Parâmetros de desempenho	Equação	
Sensibilidade (SENS)	$SENS = \frac{VP}{VP + FN}$	(Eq. 03)
Especificidade (ESPEC)	$ESPEC = \frac{VN}{FP + VN}$	(Eq. 04)
Exatidão (EXAT)	$EX = \frac{VP + VN}{VP + FN + FP + VN}$	(Eq. 05)
Taxa de Falso Positivo (TFP)	$TFP = \frac{FP}{FP + VN}$	(Eq. 06)
Taxa de Falso Negativo (TFN)	$TFN = \frac{FN}{FN + VP}$	(Eq. 07)
Precisão (PREC)	$PREC = \frac{VP}{VP + FP}$	(Eq. 08)
F1-score	$F1 = 2 \cdot \frac{(Precisão \cdot Sensibilidade)}{(Precisão + Sensibilidade)}$	(Eq. 09)
MCC	$MCC = \frac{(VP \cdot VN - FN \cdot FP)}{\sqrt{(VP + FP)(FP + FN)(VN + FP)(VN + FN)}}$	(Eq. 10)

2.5. Referências

- (1) Brereton, R. G. A Short History of Chemometrics: A Personal View. *J Chemom* **2014**, 28 (10), 749–760. <https://doi.org/10.1002/cem.2633>.
- (2) Kowalski, B. R.; Seasholtz, M. B. Recent Developments in Multivariate Calibration. *J Chemom* **1991**, 5 (3), 129–145. <https://doi.org/10.1002/cem.1180050303>.
- (3) Jurs, P. C.; Kowalski, B. R.; Isenhour, T. L. Computerized Learning Machines Applied to Chemical Problems Molecular Formula Determination from Low Resolution Mass Spectrometry. *Anal Chem* **1969**, 41 (1), 21–27. <https://doi.org/10.1021/ac60270a002>.
- (4) Jurs, P. C.; Kowalski, B. R.; Isenhour, T. L.; Reilley, C. N. Computerized Learning Machines Applied to Chemical Problems. Investigation of Convergence Rate and Predictive Ability of Adaptive Binary Pattern Classifiers. *Anal Chem* **1969**, 41 (6), 690–695. <https://doi.org/10.1021/ac60275a025>.
- (5) Kowalski, B. R.; Jurs, P. C.; Isenhour, T. L.; Reilley, C. N. Computerized Learning Machines Applied to Chemical Problems. Multicategory Pattern Classification by Least Squares. *Anal Chem* **1969**, 41 (6), 695–700. <https://doi.org/10.1021/ac60275a026>.
- (6) Kowalski, B. R.; Jurs, P. C.; Isenhour, T. L.; Reilley, C. N.; Isenhour, T. L.; Jurs, P. C. Computerized Learning Machines Applied to Chemical Problems Interpretation of

- Infrared Spectrometry Data. *Anal Chem* **1969**, *41* (14), 1945–1949. <https://doi.org/10.1021/ac50159a026>.
- (7) Kowalski, B. R.; Seasholtz, M. B. Recent Developments in Multivariate Calibration. *J Chemom* **1991**, *5* (3), 129–145. <https://doi.org/10.1002/cem.1180050303>.
 - (8) Wold, S. Pattern Recognition by Means of Disjoint Principal Components Models. *Pattern Recognit* **1976**, *8* (3), 127–139. [https://doi.org/10.1016/0031-3203\(76\)90014-5](https://doi.org/10.1016/0031-3203(76)90014-5).
 - (9) Hopke, P. K. The Evolution of Chemometrics. *Anal Chim Acta* **2003**, *500* (1–2), 365–377. [https://doi.org/10.1016/S0003-2670\(03\)00944-9](https://doi.org/10.1016/S0003-2670(03)00944-9).
 - (10) Kowalski, B. R. Chemometrics: Views and Propositions? *J Chem Inf Comput Sci* **1975**, *15* (4), 201–203.
 - (11) Ferreira, M. M. C. *Quimiometria - Conceitos, Métodos e Aplicações*, 1st ed.; Editora da UNICAMP, Ed.; Campinas, 2015.
 - (12) Março, P. H.; Valderrama, P.; Alexandrino, G. L.; Poppi, R. J.; Tauler, R. Resolução Multivariada de Curvas Com Mínimos Quadrados Alternantes: Descrição, Funcionamento e Aplicações. *Quim Nova* **2014**, *37* (9), 1525–1532. <https://doi.org/10.5935/0100-4042.20140205>.
 - (13) Bro, R. *Chemometrics and Intelligent Laboratory Systems Tutorial PARAFAC. Tutorial and Applications*; 1997; Vol. 38.
 - (14) Breiman, L. Random Forests. *Mach Learn* **2001**, *45* (1), 5–32. <https://doi.org/10.1023/A:101093340>.
 - (15) Lovatti, B. P. O.; Nascimento, M. H. C.; Neto, Á. C.; Castro, E. V. R.; Filgueiras, P. R. Use of Random Forest in the Identification of Important Variables. *Microchemical Journal* **2019**, *145*, 1129–1134. <https://doi.org/10.1016/j.microc.2018.12.028>.
 - (16) Oliveira, B. P. L.; Nascimento, M. H. C.; Rainha, K. P.; Alvaro, E. C. S. O.; Filgueiras, P. R. Different Strategies for the Use of Random Forest in NMR Spectra. *J. Chemometrics* **2020**, No. Special issue, 1–10. <https://doi.org/10.1002/cem.3231>.
 - (17) de Santana, F. B.; Borges Neto, W.; Poppi, R. J. Random Forest as One-Class Classifier and Infrared Spectroscopy for Food Adulteration Detection. *Food Chem* **2019**, *293*, 323–332. <https://doi.org/10.1016/j.foodchem.2019.04.073>.
 - (18) de Santana, F. B.; de Souza, A. M.; Poppi, R. J. Visible and near Infrared Spectroscopy Coupled to Random Forest to Quantify Some Soil Quality Parameters. *Spectrochim Acta A Mol Biomol Spectrosc* **2018**, *191*, 454–462. <https://doi.org/10.1016/j.saa.2017.10.052>.
 - (19) Guleken, Z.; Jakubczyk, P.; Wiesław, P.; Krzysztof, P.; Bulut, H.; Öten, E.; Depciuch, J.; Tarhan, N. Characterization of Covid-19 Infected Pregnant Women Sera Using Laboratory Indexes, Vibrational Spectroscopy, and Machine Learning Classifications. *Talanta* **2022**, *237*. <https://doi.org/10.1016/j.talanta.2021.122916>.
 - (20) Nascimento, M. H. C.; Marcarini, W. D.; Folli, G. S.; da Silva Filho, W. G.; Barbosa, L. L.; Paulo, E. H. de; Vassallo, P. F.; Mill, J. G.; Barauna, V. G.; Martin, F. L.; de Castro, E. V. R.; Romão, W.; Filgueiras, P. R. Noninvasive Diagnostic for COVID-19

from Saliva Biofluid via FTIR Spectroscopy and Multivariate Analysis. *Anal Chem* **2022**, *94* (5), 2425–2433. <https://doi.org/10.1021/acs.analchem.1c04162>.

- (21) Afanador, N. L.; Smolinska, A.; Tran, T. N.; Blanchet, L. Unsupervised Random Forest: A Tutorial with Case Studies. *J. Chemometrics* **2016**, *30* (5), 232–241. <https://doi.org/10.1002/cem.2790>.
- (22) Nogueira, M. S.; Leal, L. B.; Macarini, W.; Pimentel, R. L.; Muller, M.; Vassallo, P. F.; Campos, L. C. G.; dos Santos, L.; Luiz, W. B.; Mill, J. G.; Barauna, V. G.; de Carvalho, L. F. das C. e. S. Rapid Diagnosis of COVID-19 Using FT-IR ATR Spectroscopy and Machine Learning. *Sci Rep* **2021**, *11* (1). <https://doi.org/10.1038/s41598-021-93511-2>.
- (23) Callery, E. L.; Rowbottom, A. W. Vibrational Spectroscopy and Multivariate Analysis Techniques in the Clinical Immunology Laboratory: A Review of Current Applications and Requirements for Diagnostic Use. **2021**. <https://doi.org/10.1080/05704928.2021.1958337?needAccess=true>.
- (24) Cameron, J. M.; Brennan, P. M.; Antoniou, G.; Butler, H. J.; Christie, L.; Conn, J. J. A.; Curran, T.; Gray, E.; Hegarty, M. G.; Jenkinson, M. D.; Orringer, D.; Palmer, D. S.; Sala, A.; Smith, B. R.; Baker, M. J. Clinical Validation of a Spectroscopic Liquid Biopsy for Earlier Detection of Brain Cancer. *Neurooncol Adv* **2022**, *4* (1). <https://doi.org/10.1093/oaajnl/vdac024>.
- (25) Garza, K. Y.; Silva, A. A. R.; Rosa, J. R.; Keating, M. F.; Povilaitis, S. C.; Spradlin, M.; Sanches, P. H. G.; Varão Moura, A.; Marrero Gutierrez, J.; Lin, J. Q.; Zhang, J.; DeHoog, R. J.; Bensussan, A.; Badal, S.; Cardoso de Oliveira, D.; Dias Garcia, P. H.; Dias de Oliveira Negrini, L.; Antonio, M. A.; Canevari, T. C.; Eberlin, M. N.; Tibshirani, R.; Eberlin, L. S.; Porcari, A. M. Rapid Screening of COVID-19 Directly from Clinical Nasopharyngeal Swabs Using the MasSpec Pen. *Anal Chem* **2021**, *93* (37), 12582–12593. <https://doi.org/10.1021/acs.analchem.1c01937>.
- (26) Santos, M. C. D.; Morais, C. L. M.; Lima, K. M. G. ATR-FTIR Spectroscopy for Virus Identification: A Powerful Alternative. *Biomed Spectrosc Imaging* **2020**, *9* (3–4), 103–118. <https://doi.org/10.3233/bsi-200203>.
- (27) Naseer, K.; Ali, S.; Qazi, J. ATR-FTIR Spectroscopy as the Future of Diagnostics: A Systematic Review of the Approach Using Bio-Fluids. *Applied Spectroscopy Reviews*. Bellwether Publishing, Ltd. 2021, pp 85–97. <https://doi.org/10.1080/05704928.2020.1738453>.
- (28) Finlayson, D.; Rinaldi, C.; Baker, M. J. Is Infrared Spectroscopy Ready for the Clinic? *Anal. Chem.* **2019**, *91* (19), 12117–12128. <https://doi.org/10.1021/acs.analchem.9b02280>.
- (29) Olivieri, A. C. Analytical Advantages of Multivariate Data Processing. One, Two, Three, Infinity? *Anal Chem* **2008**, *80* (15), 5713–5720. <https://doi.org/10.1021/ac800692c>.
- (30) Baker, M. J.; Trevisan, J.; Bassan, P.; Bhargava, R.; Butler, H. J.; Dorling, K. M.; Fielden, P. R.; Fogarty, S. W.; Fullwood, N. J.; Heys, K. A.; Hughes, C.; Lasch, P.; Martin-Hirsch, P. L.; Obinaju, B.; Sockalingum, G. D.; Sulé-Suso, J.; Strong, R. J.; Walsh, M. J.; Wood, B. R.; Gardner, P.; Martin, F. L. Using Fourier Transform IR

Spectroscopy to Analyze Biological Materials. *Nat. Protoc.* **2014**, 9 (8), 1771–1791. <https://doi.org/10.1038/nprot.2014.110>.

- (31) Banerjee, A.; Gokhale, A.; Bankar, R.; Palanivel, V.; Salkar, A.; Robinson, H.; Shastri, J. S.; Agrawal, S.; Hartel, G.; Hill, M. M.; Srivastava, S. Rapid Classification of COVID-19 Severity by ATR-FTIR Spectroscopy of Plasma Samples. *Anal. Chem.* **2021**, 93 (30), 10391–10396. <https://doi.org/10.1021/acs.analchem.1c00596>.
- (32) Morais, C. L. M.; Paraskevaidi, M.; Cui, L.; Fullwood, N. J.; Isabelle, M.; Lima, K. M. G.; Martin-Hirsch, P. L.; Sreedhar, H.; Trevisan, J.; Walsh, M. J.; Zhang, D.; Zhu, Y. G.; Martin, F. L. Standardization of Complex Biologically Derived Spectrochemical Datasets. *Nat. Protoc.* **2019**, 14 (5), 1546–1577. <https://doi.org/10.1038/s41596-019-0150-x>.
- (33) Rutherford, S. H.; Nordon, A.; Hunt, N. T.; Baker, M. J. Biofluid Analysis and Classification Using IR and 2D-IR Spectroscopy. *Chemometrics Intell. Lab. Syst.* **2021**, 217 (104408), 1–11. <https://doi.org/10.1016/j.chemolab.2021.104408>.
- (34) Barauna, V. G.; Singh, M. N.; Barbosa, L. L.; Marcarini, W. D.; Vassallo, P. F.; Mill, J. G.; Ribeiro-Rodrigues, R.; Campos, L. C. G.; Warnke, P. H.; Martin, F. L. Ultrarapid On-Site Detection of SARS-CoV-2 Infection Using Simple ATR-FTIR Spectroscopy and an Analysis Algorithm: High Sensitivity and Specificity. *Anal. Chem.* **2021**, 93 (5), 2950–2958. <https://doi.org/10.1021/acs.analchem.0c04608>.
- (35) Zhou, Z.-H. Ensemble Methods: Foundations and Algorithms. In *Ensemble Methods: Foundations and Algorithms*; Taylos & Francis Group, Ed.; 2012; pp 15–20.
- (36) Khoury, E. Y.; Gebelin, M.; de Sèze, J.; Patte-Mensah, C.; Marcou, G.; Varnek, A.; Mensah-Nyagan, A. G.; Hellwig, P.; Collongues, N. Rapid Discrimination of Neuromyelitis Optica Spectrum Disorder and Multiple Sclerosis Using Machine Learning on Infrared Spectra of Sera. *Int J Mol Sci* **2022**, 23 (5). <https://doi.org/10.3390/ijms23052791>.
- (37) Moisoiu, T.; Iancu, S. D.; Burghilea, D.; Dragomir, M. P.; Iacob, G.; Stefancu, A.; Cozan, R. G.; Antal, O.; Bálint, Z.; Muntean, V.; Badea, R. I.; Licarete, E.; Leopold, N.; Elec, F. I. SERS Liquid Biopsy Profiling of Serum for the Diagnosis of Kidney Cancer. *Biomedicines* **2022**, 10 (2). <https://doi.org/10.3390/biomedicines10020233>.
- (38) Massimo, A.; Cuccurullo, C. Bibliometrix: An R-Tool for Comprehensive Science Mapping Analysis. *J Informetr* **2017**, 11, 959–975.
- (39) Pagani, R. N.; Kovalski, J. L.; Resende, L. M. Methodi Ordinatio: A Proposed Methodology to Select and Rank Relevant Scientific Papers Encompassing the Impact Factor, Number of Citation, and Year of Publication. *Scientometrics* **2015**, 105, 2109–2135. <https://doi.org/10.1007/s11192-015-1744-x>.
- (40) Esbensen, K.; Geladi, P. The Start and Early History of Chemometrics: Selected Interviews. Part 2. *J Chemom* **1990**, 4 (6), 389–412. <https://doi.org/10.1002/cem.1180040604>.
- (41) Brown, P. S. Some Perspectives on the History and Sociology of the Chemometrics Revolution - and Some Suggestions for What the Future Holds. In *Chemometrics Conference*; França, 2019.

- (42) Pearson, K. On Lines and Planes of Closest Fit to Systems of Points in Space. *Philosophical Magazine Series* **1901**, 6 (2(11)), 559–572.
- (43) Hotelling, H. Analysis of a Complex of Statistical Variables Into Principal Components. *Warwick and York* **1933**.
- (44) Tharwat, A. Principal Component Analysis - a Tutorial. *International Journal of Applied Pattern Recognition* **2016**, 3 (3), 197. <https://doi.org/10.1504/ijapr.2016.10000630>.
- (45) Fukunaga, K.; Koontz, W. L. G. *Application of the Karhunen-Loeve Expansion to Feature Selection and Ordering*; 1970; Vol. 19.
- (46) Bruns, R. E. Como Nasceu a Quimiometria No Brasil. In *Workshop de Quimiometria*; Campina Grande, 2020.
- (47) Fisher, R. A. The Use of Multiple Measurements in Taxonomic Problems. *Ann Eugen* **1936**, 7 (2), 179–188. <https://doi.org/10.1111/j.1469-1809.1936.tb02137.x>.
- (48) Fisher, R. A.; Mackenzie, W. A. The Manurial Response of Diferent Potato Varieties. *Journal of agricultural Science*. 1923, pp 311–320.
- (49) Jurs, P. C.; Kowalski, B. R.; Isenhour, T. L.; Reilley, C. N. An Investigation of Combined Patterns from Diverse Analytical Data Using Computerized Learning Machines. *Anal Chem* **1969**, 41 (14), 1949–1953. <https://doi.org/10.1021/ac50159a027>.
- (50) Kowalski, B. R.; Bender, C. F. Pattern Recognition.1 A Powerful Approach to Interpreting Chemical Data. *J Am Chem Soc* **1972**, 94 (16), 5632–5639. <https://doi.org/10.1021/ja00771a016>.
- (51) Lawton, W. H.; Sylvestre, E. A. Self Modeling Curve Resolution. *Technometrics* **1971**, 13 (3), 617–633. <https://doi.org/10.1080/00401706.1971.10488823>.
- (52) Kiralj, R.; Ferreira, M. M. C. The Past, Present, Adn Future of Chemometrics Worldwide: Some Etymological, Linguistic, and Bibliometric Investigations. *J Chemom* **2006**, 20, 247–272. <https://doi.org/10.1002/cem.1001>.
- (53) Barros Neto, B. de; Scarminio, I. S.; Bruns, R. E. 25 Anos de Quimiometria No Brasil. *Quim Nova* **2006**, 29 (6), 1401–1406. <https://doi.org/10.1590/s0100-40422006000600042>.
- (54) Sena, M. M.; Braga, J. W. B.; Scarminio, I. S.; Bruns, R. E. A Tribute to Professor Ronei J. Poppi, a Pioneer of Multivariate Calibration in South America and a Prolific Mentor of Chemometricians in Brazil. *Journal of Chemometrics*. John Wiley and Sons Ltd December 1, 2020. <https://doi.org/10.1002/cem.3284>.
- (55) Special Issue Dedicated to Prof. Roney J. Poppi. *Brazilian Journal of Analytical Chemistry* **2021**, 8 (32), 1–270.
- (56) Breitzkreitz, M. C. Analytical Quality by Design. *Brazilian Journal of Analytical Chemistry*. DKK Comunicacao 2021, pp 1–5. <https://doi.org/10.30744/BRJAC.2179-3425.EDITORIAL.MCBREITKREITZ.N32>.

- (57) Hibbert, D. B. Vocabulary of Concepts and Terms in Chemometrics (IUPAC Recommendations 2016). *Pure and Applied Chemistry* **2016**, 88 (4), 407–443. <https://doi.org/10.1515/pac-2015-0605>.
- (58) Brereton, R. G. Pattern Recognition in Chemometrics. *Chemometrics and Intelligent Laboratory Systems* **2015**, 149, 90–96. <https://doi.org/10.1016/j.chemolab.2015.06.012>.
- (59) Amigo, J. M. Data Mining, Machine Learning, Deep Learning, Chemometrics: Definitions, Common Points and Trends (Spoiler Alert: VALIDATE Your Models!). *Brazilian Journal of Analytical Chemistry* **2021**, 8 (32), 22–38. <https://doi.org/10.30744/BRJAC.2179-3425.AR-38-2021>.
- (60) Mathworks. *Machine Learning with Matlab*. <https://www.mathworks.com/content/dam/mathworks/ebook/gated/machine-learning-ebook-all-chapters.pdf> (accessed 2021-01-26).
- (61) Brereton, R. G.; Lloyd, G. R. Support Vector Machines for Classification and Regression. *Analyst*. Royal Society of Chemistry 2010, pp 230–267. <https://doi.org/10.1039/b918972f>.
- (62) Kwon, S. J. Artificial Neural Networks. *Artificial Neural Networks* **2011**, 1–426. <https://doi.org/10.15864/jmscm.1104>.
- (63) Izbicki, R.; Santos dos, T. M. *Aprendizado de Máquina: Uma Abordagem Estatística*; São Carlos, 2020; Vol. 1.
- (64) Jiménez-Carvelo, A. M.; González-Casado, A.; Bagur-González, M. G.; Cuadros-Rodríguez, L. Alternative Data Mining/Machine Learning Methods for the Analytical Evaluation of Food Quality and Authenticity – A Review. *Food Research International*. Elsevier Ltd August 1, 2019, pp 25–39. <https://doi.org/10.1016/j.foodres.2019.03.063>.
- (65) Mitchell B.O., J. B. O. Machine Learning Methods in Chemoinformatics. *Wiley Interdisciplinary Reviews: Computational Molecular Science*. Blackwell Publishing Inc. 2014, pp 468–481. <https://doi.org/10.1002/wcms.1183>.
- (66) Barea-Sepúlveda, M.; Ferreiro-González, M.; Calle, J. L. P.; Barbero, G. F.; Ayuso, J.; Palma, M. Comparison of Different Processing Approaches by SVM and RF on HS-MS ENose and NIR Spectrometry Data for the Discrimination of Gasoline Samples. *Microchemical Journal* **2022**, 172. <https://doi.org/10.1016/j.microc.2021.106893>.
- (67) Ho, K. The Random Subspace Method for Constructing Decision Forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **1998**, 20 (8), 832–844.
- (68) Quinlan, J. R. Induction of Decision Trees. *Mach Learn* **1986**, 1, 81–106.
- (69) Breiman, L. Bagging Predictors. *Mach. Learn.* **1996**, 24 (2), 123–140. <https://doi.org/10.1007/bf00058655>.
- (70) Ho, K.; Hull, J. J.; Srihari, S. N. *Decision Combination in Multiple Classifier Systems*; 1994; Vol. 16.
- (71) Breiman, L.; Friedman, J. H.; Olshen, R. A.; Stone, C. J. *Classification and Regression Trees*, 1st ed.; Taylor & Francis, 1984; Vol. 1. <https://doi.org/10.1201/9781315139470>.

- (72) Lee, S.; Choi, H.; Cha, K.; Chung, H. Random Forest as a Potential Multivariate Method for Near-Infrared (NIR) Spectroscopic Analysis of Complex Mixture Samples: Gasoline and Naphtha. *Microchemical Journal* **2013**, *110*, 739–748. <https://doi.org/10.1016/j.microc.2013.08.007>.
- (73) de Santana, F. B.; Mazivila, S. J.; Gontijo, L. C.; Neto, W. B.; Poppi, R. J. Rapid Discrimination Between Authentic and Adulterated Andiroba Oil Using FTIR-HATR Spectroscopy and Random Forest. *Food Anal Methods* **2018**, *11* (7), 1927–1935. <https://doi.org/10.1007/s12161-017-1142-5>.
- (74) Wu, X. M.; Zhang, Q. Z.; Wang, Y. Z. Traceability of Wild Paris Polyphylla Smith Var. Yunnanensis Based on Data Fusion Strategy of FT-MIR and UV–Vis Combined with SVM and Random Forest. *Spectrochim Acta A Mol Biomol Spectrosc* **2018**, *205*, 479–488. <https://doi.org/10.1016/j.saa.2018.07.067>.
- (75) Svetnik, V.; Liaw, A.; Tong, C.; Wang, T. LNCS 3077 - Application of Breiman's Random Forest to Modeling Structure-Activity Relationships of Pharmaceutical Molecules. In *MCS: International Workshop on Multiple Classifier Systems*; F. Roli, J. Kittler, and T. Windeatt: Berlin, 2004; pp 334–343.
- (76) Svetnik, V.; Liaw, A.; Tong, C.; Christopher Culberson, J.; Sheridan, R. P.; Feuston, B. P. Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling. *J Chem Inf Comput Sci* **2003**, *43* (6), 1947–1958. <https://doi.org/10.1021/ci034160g>.
- (77) Dotto, A. C.; Dalmolin, R. S. D.; ten Caten, A.; Grunwald, S. A Systematic Study on the Application of Scatter-Corrective and Spectral-Derivative Preprocessing for Multivariate Prediction of Soil Organic Carbon by Vis-NIR Spectra. *Geoderma* **2018**, *314*, 262–274. <https://doi.org/10.1016/j.geoderma.2017.11.006>.
- (78) Hounkpatin, O. K. L.; op de Hipt, F.; Bossa, A. Y.; Welp, G.; Amelung, W. Soil Organic Carbon Stocks and Their Determining Factors in the Dano Catchment (Southwest Burkina Faso). *Catena (Amst)* **2018**, *166*, 298–309. <https://doi.org/10.1016/j.catena.2018.04.013>.
- (79) Hopley, E.; Steffens, M.; Bauke, S. L.; Kögel-Knabner, I. Hotspots of Soil Organic Carbon Storage Revealed by Laboratory Hyperspectral Imaging. *Sci Rep* **2018**, *8* (1). <https://doi.org/10.1038/s41598-018-31776-w>.
- (80) Yaseen, Z. M. An Insight into Machine Learning Models Era in Simulating Soil, Water Bodies and Adsorption Heavy Metals: Review, Challenges and Solutions. *Chemosphere* **2021**, *277*. <https://doi.org/10.1016/j.chemosphere.2021.130126>.
- (81) Gislason, P. O.; Benediktsson, J. A.; Sveinsson, J. R. Random Forests for Land Cover Classification. In *Pattern Recognition Letters*; 2006; Vol. 27, pp 294–300. <https://doi.org/10.1016/j.patrec.2005.08.011>.
- (82) Wu, B.; Abbott, T.; Fishman, D.; McMurray, W.; Mor, G.; Stone, K.; Ward, D.; Williams, K.; Zhao, H. Comparison of Statistical Methods for Classification of Ovarian Cancer Using Mass Spectrometry Data. *Bioinformatics* **2003**, *19* (13), 1636–1643. <https://doi.org/10.1093/bioinformatics/btg210>.
- (83) Kalogiouri, N. P.; Aalizadeh, R.; Dasenaki, M. E.; Thomaidis, N. S. Application of High Resolution Mass Spectrometric Methods Coupled with Chemometric Techniques in

Olive Oil Authenticity Studies - A Review. *Analytica Chimica Acta*. Elsevier B.V. October 16, 2020, pp 150–173. <https://doi.org/10.1016/j.aca.2020.07.029>.

- (84) Nawar, S.; Delbecq, N.; Declercq, Y.; de Smedt, P.; Finke, P.; Verdoodt, A.; van Meirvenne, M.; Mouazen, A. M. Can Spectral Analyses Improve Measurement of Key Soil Fertility Parameters with X-Ray Fluorescence Spectrometry? *Geoderma* **2019**, *350*, 29–39. <https://doi.org/10.1016/j.geoderma.2019.05.002>.
- (85) Tian, Y.; Yan, C.; Zhang, T.; Tang, H.; Li, H.; Yu, J.; Bernard, J.; Chen, L.; Martin, S.; Delepine-Gilon, N.; Bocková, J.; Veis, P.; Chen, Y.; Yu, J. Classification of Wines According to Their Production Regions with the Contained Trace Elements Using Laser-Induced Breakdown Spectroscopy. *Spectrochim Acta Part B At Spectrosc* **2017**, *135*, 91–101. <https://doi.org/10.1016/j.sab.2017.07.003>.
- (86) Shen, T.; Kong, W.; Liu, F.; Chen, Z.; Yao, J.; Wang, W.; Peng, J.; Chen, H.; He, Y. Rapid Determination of Cadmium Contamination in Lettuce Using Laser-Induced Breakdown Spectroscopy. *Molecules* **2018**, *23* (11). <https://doi.org/10.3390/molecules23112930>.
- (87) Stefan, D.; Gyftokostas, N.; Kourelas, P.; Nanou, E.; Tananaki, C.; Kanelis, D.; Liolios, V.; Kokkinos, V.; Bouras, C.; Couris, S. Honey Discrimination Based on the Bee Feeding by Laser Induced Breakdown Spectroscopy. *Food Control* **2022**, *134*. <https://doi.org/10.1016/j.foodcont.2021.108770>.
- (88) Sun, H.; Song, C.; Lin, X.; Gao, X. Identification of Meat Species by Combined Laser-Induced Breakdown and Raman Spectroscopies. *Spectrochim Acta Part B At Spectrosc* **2022**, *194*. <https://doi.org/10.1016/j.sab.2022.106456>.
- (89) Qi, J.; Zhang, T.; Tang, H.; Li, H. Rapid Classification of Archaeological Ceramics via Laser-Induced Breakdown Spectroscopy Coupled with Random Forest. *Spectrochim Acta Part B At Spectrosc* **2018**, *149*, 288–293. <https://doi.org/10.1016/j.sab.2018.09.006>.
- (90) Rzecki, K.; Sośnicki, T.; Baran, M.; Niedźwiecki, M.; Król, M.; Łojewski, T.; Acharya, U. R.; Yildirim, Ö.; Pławiak, P. Application of Computational Intelligence Methods for the Automated Identification of Paper-Ink Samples Based on LIBS. *Sensors* **2018**, *18* (11). <https://doi.org/10.3390/S18113670>.
- (91) Chu, Y.; Chen, T.; Chen, F.; Tang, Y.; Tang, S.; Jin, H.; Guo, L.; Lu, Y. F.; Zeng, X. Discrimination of Nasopharyngeal Carcinoma Serum Using Laser-Induced Breakdown Spectroscopy Combined with an Extreme Learning Machine and Random Forest Method. *J Anal At Spectrom* **2018**, *33* (12), 2083–2088. <https://doi.org/10.1039/c8ja00263k>.
- (92) Liu, K.; Tian, D.; Wang, H.; Yang, G. Rapid Classification of Plastics by Laser-Induced Breakdown Spectroscopy (LIBS) Coupled with Partial Least Squares Discrimination Analysis Based on Variable Importance (VI-PLS-DA). *Analytical Methods* **2019**, *11* (9), 1214–1223. <https://doi.org/10.1039/c8ay02755b>.
- (93) Menze, B. H.; Petrich, W.; Hamprecht, F. A. Multivariate Feature Selection and Hierarchical Classification for Infrared Spectroscopy: Serum-Based Detection of Bovine Spongiform Encephalopathy. In *Analytical and Bioanalytical Chemistry*; 2007; Vol. 387, pp 1801–1807. <https://doi.org/10.1007/s00216-006-1070-5>.

- (94) Wang, J.; Ding, J.; Abulimiti, A.; Cai, L. Quantitative Estimation of Soil Salinity by Means of Different Modeling Methods and Visible-near Infrared (VIS-NIR) Spectroscopy, Ebinur Lake Wetland, Northwest China. *PeerJ* **2018**, *2018* (5). <https://doi.org/10.7717/peerj.4703>.
- (95) Liu, W.; Liu, C.; Yu, J.; Zhang, Y.; Li, J.; Chen, Y.; Zheng, L. Discrimination of Geographical Origin of Extra Virgin Olive Oils Using Terahertz Spectroscopy Combined with Chemometrics. *Food Chem* **2018**, *251*, 86–92. <https://doi.org/10.1016/j.foodchem.2018.01.081>.
- (96) Liu, W.; Zhao, P.; Wu, C.; Liu, C.; Yang, J.; Zheng, L. Rapid Determination of Aflatoxin B1 Concentration in Soybean Oil Using Terahertz Spectroscopy with Chemometric Methods. *Food Chem* **2019**, *293*, 213–219. <https://doi.org/10.1016/j.foodchem.2019.04.081>.
- (97) Estelles-Lopez, L.; Ropodi, A.; Pavlidis, D.; Fotopoulou, J.; Gkousari, C.; Peyrodie, A.; Panagou, E.; Nychas, G. J.; Mohareb, F. An Automated Ranking Platform for Machine Learning Regression Models for Meat Spoilage Prediction Using Multi-Spectral Imaging and Metabolic Profiling. *Food Research International* **2017**, *99*, 206–215. <https://doi.org/10.1016/j.foodres.2017.05.013>.
- (98) Ropodi, A. I.; Panagou, E. Z.; Nychas, G. J. E. Multispectral Imaging (MSI): A Promising Method for the Detection of Minced Beef Adulteration with Horsemeat. *Food Control* **2017**, *73*, 57–63. <https://doi.org/10.1016/j.foodcont.2016.05.048>.
- (99) Weng, S.; Wang, F.; Dong, R.; Qiu, M.; Zhao, J.; Huang, L.; Zhang, D. Fast and Quantitative Analysis of Ediphenphos Residue in Rice Using Surface-Enhanced Raman Spectroscopy. *J Food Sci* **2018**, *83* (4), 1179–1185. <https://doi.org/10.1111/1750-3841.14103>.
- (100) Weng, S.; Guo, B.; Tang, P.; Yin, X.; Pan, F.; Zhao, J.; Huang, L.; Zhang, D. Rapid Detection of Adulteration of Minced Beef Using Vis/NIR Reflectance Spectroscopy with Multivariate Methods. *Spectrochim Acta A Mol Biomol Spectrosc* **2020**, *230*. <https://doi.org/10.1016/j.saa.2019.118005>.
- (101) Hart, C. D.; Vignoli, A.; Tenori, L.; Uy, G. L.; van To, T.; Adebamowo, C.; Hossain, S. M.; Biganzoli, L.; Risi, E.; Love, R. R.; Luchinat, C.; di Leo, A. Serum Metabolomic Profiles Identify ER-Positive Early Breast Cancer Patients at Increased Risk of Disease Recurrence in a Multicenter Population. *Clinical Cancer Research* **2017**, *23* (6), 1422–1431. <https://doi.org/10.1158/1078-0432.CCR-16-1153>.
- (102) Gunther, E. C.; Stone, D. J.; Gerwien, R. W.; Bento, P.; Heyes, M. P. Prediction of Clinical Drug Efficacy by Classification of Drug-Induced Genomic Expression Profiles in Vitro. *PNAS* **2003**, *100* (16), 9608–9613. <https://doi.org/doi-org.ez43>.
- (103) Díaz-Uriarte, R.; Alvarez de Andrés, S. Gene Selection and Classification of Microarray Data Using Random Forest. *BMC Bioinformatics* **2006**, *7*. <https://doi.org/10.1186/1471-2105-7-3>.
- (104) Cheng, X.; Vella, A.; Stasiewicz, M. J. Classification of Aflatoxin Contaminated Single Corn Kernels by Ultraviolet to near Infrared Spectroscopy. *Food Control* **2019**, *98*, 253–261. <https://doi.org/10.1016/j.foodcont.2018.11.037>.

- (105) Gold, K. M.; Townsend, P. A.; Herrmann, I.; Gevens, A. J. Investigating Potato Late Blight Physiological Differences across Potato Cultivars with Spectroscopy and Machine Learning. *Plant Science* **2020**, 295, 110316. <https://doi.org/10.1016/j.plantsci.2019.110316>.
- (106) Harrison, D.; Rivard, B.; Sánchez-Azofeifa, A. Classification of Tree Species Based on Longwave Hyperspectral Data from Leaves, a Case Study for a Tropical Dry Forest. *International Journal of Applied Earth Observation and Geoinformation* **2018**, 66, 93–105. <https://doi.org/10.1016/j.jag.2017.11.009>.
- (107) Moura-Bueno, J. M.; Dalmolin, R. S. D.; ten Caten, A.; Dotto, A. C.; Demattê, J. A. M. Stratification of a Local VIS-NIR-SWIR Spectral Library by Homogeneity Criteria Yields More Accurate Soil Organic Carbon Predictions. *Geoderma* **2019**, 337, 565–581. <https://doi.org/10.1016/j.geoderma.2018.10.015>.
- (108) Nawar, S.; Mouazen, A. M. On-Line Vis-NIR Spectroscopy Prediction of Soil Organic Carbon Using Machine Learning. *Soil Tillage Res* **2019**, 190, 120–127. <https://doi.org/10.1016/j.still.2019.03.006>.
- (109) Gholizadeh, A.; Saberioon, M.; Viscarra Rossel, R. A.; Boruvka, L.; Klement, A. Spectroscopic Measurements and Imaging of Soil Colour for Field Scale Estimation of Soil Organic Carbon. *Geoderma* **2020**, 357, 113972. <https://doi.org/10.1016/j.geoderma.2019.113972>.
- (110) Bao, Y.; Meng, X.; Ustin, S.; Wang, X.; Zhang, X.; Liu, H.; Tang, H. Vis-SWIR Spectral Prediction Model for Soil Organic Matter with Different Grouping Strategies. *Catena (Amst)* **2020**, 195, 104703. <https://doi.org/10.1016/j.catena.2020.104703>.
- (111) Wang, J.; Ding, J.; Yu, D.; Ma, X.; Zhang, Z.; Ge, X.; Teng, D.; Li, X.; Liang, J.; Lizaga, I.; Chen, X.; Yuan, L.; Guo, Y. Capability of Sentinel-2 MSI Data for Monitoring and Mapping of Soil Salinity in Dry and Wet Seasons in the Ebinur Lake Region, Xinjiang, China. *Geoderma* **2019**, 353, 172–187. <https://doi.org/10.1016/j.geoderma.2019.06.040>.
- (112) Benedet, L.; Faria, W. M.; Silva, S. H. G.; Mancini, M.; Demattê, J. A. M.; Guilherme, L. R. G.; Curi, N. Soil Texture Prediction Using Portable X-Ray Fluorescence Spectrometry and Visible near-Infrared Diffuse Reflectance Spectroscopy. *Geoderma* **2020**, 376, 114553. <https://doi.org/10.1016/j.geoderma.2020.114553>.
- (113) Hong, Y.; Shen, R.; Cheng, H.; Chen, Y.; Zhang, Y.; Liu, Y.; Zhou, M.; Yu, L.; Liu, Y.; Liu, Y. Estimating Lead and Zinc Concentrations in Peri-Urban Agricultural Soils through Reflectance Spectroscopy: Effects of Fractional-Order Derivative and Random Forest. *Science of The Total Environment* **2019**, 651, 1969–1982. <https://doi.org/10.1016/j.scitotenv.2018.09.391>.
- (114) Zhou, W.; Yang, H.; Xie, L.; Li, H.; Huang, L.; Zhao, Y.; Yue, T. Hyperspectral Inversion of Soil Heavy Metals in Three-River Source Region Based on Random Forest Model. *Catena (Amst)* **2021**, 202, 105222. <https://doi.org/10.1016/j.catena.2021.105222>.
- (115) Hong, Y.; Shen, R.; Cheng, H.; Chen, S.; Chen, Y.; Guo, L.; He, J.; Liu, Y.; Yu, L.; Liu, Y. Cadmium Concentration Estimation in Peri-Urban Agricultural Soils: Using Reflectance Spectroscopy, Soil Auxiliary Information, or a Combination of Both? *Geoderma* **2019**, 354, 113875. <https://doi.org/10.1016/j.geoderma.2019.07.033>.

- (116) Tan, K.; Ma, W.; Wu, F.; Du, Q. Random Forest–Based Estimation of Heavy Metal Concentration in Agricultural Soils with Hyperspectral Sensor Data. *Environ Monit Assess* **2019**, *191* (7), 446. <https://doi.org/10.1007/s10661-019-7510-4>.
- (117) Douglas, R. K.; Nawar, S.; Alamar, M. C.; Mouazen, A. M.; Coulon, F. Rapid Prediction of Total Petroleum Hydrocarbons Concentration in Contaminated Soil Using Vis-NIR Spectroscopy and Regression Techniques. *Science of The Total Environment* **2018**, *616–617*, 147–155. <https://doi.org/10.1016/j.scitotenv.2017.10.323>.
- (118) Douglas, R. K.; Nawar, S.; Cipullo, S.; Alamar, M. C.; Coulon, F.; Mouazen, A. M. Evaluation of Vis-NIR Reflectance Spectroscopy Sensitivity to Weathering for Enhanced Assessment of Oil Contaminated Soils. *Science of The Total Environment* **2018**, *626*, 1108–1120. <https://doi.org/10.1016/j.scitotenv.2018.01.122>.
- (119) Wang, L.; Chang, Q.; Yang, J.; Zhang, X.; Li, F. Estimation of Paddy Rice Leaf Area Index Using Machine Learning Methods Based on Hyperspectral Data from Multi-Year Experiments. *PLoS One* **2018**, *13* (12), e0207624. <https://doi.org/10.1371/journal.pone.0207624>.
- (120) Torrisi, S. B.; Carbone, M. R.; Rohr, B. A.; Montoya, J. H.; Ha, Y.; Yano, J.; Suram, S. K.; Hung, L. Random Forest Machine Learning Models for Interpretable X-Ray Absorption near-Edge Structure Spectrum-Property Relationships. *NPJ Comput Mater* **2020**, *6* (1), 109. <https://doi.org/10.1038/s41524-020-00376-6>.
- (121) Vinay Kumar, B. N.; Löschel, L. A.; Imhof, H. K.; Löder, M. G. J.; Laforsch, C. Analysis of Microplastics of a Broad Size Range in Commercially Important Mussels by Combining FTIR and Raman Spectroscopy Approaches. *Environmental Pollution* **2021**, *269*, 116147. <https://doi.org/10.1016/j.envpol.2020.116147>.
- (122) Li, Y.; Wang, Y. Synergistic Strategy for the Geographical Traceability of Wild Boletus Tomentipes by Means of Data Fusion Analysis. *Microchemical Journal* **2018**, *140*, 38–46. <https://doi.org/10.1016/j.microc.2018.04.001>.
- (123) Cameron, J. M.; Butler, H. J.; Smith, B. R.; Hegarty, M. G.; Jenkinson, M. D.; Syed, K.; Brennan, P. M.; Ashton, K.; Dawson, T.; Palmer, D. S.; Baker, M. J. Developing Infrared Spectroscopic Detection for Stratifying Brain Tumour Patients: Glioblastoma Multiforme vs. Lymphoma. *Analyst* **2019**, *144* (22), 6736–6750. <https://doi.org/10.1039/C9AN01731C>.
- (124) Li, Y.; Zhang, J.-Y.; Wang, Y.-Z. FT-MIR and NIR Spectral Data Fusion: A Synergetic Strategy for the Geographical Traceability of Panax Notoginseng. *Anal Bioanal Chem* **2018**, *410* (1), 91–103. <https://doi.org/10.1007/s00216-017-0692-0>.
- (125) Zhou, Y.; Zuo, Z.; Xu, F.; Wang, Y. Origin Identification of Panax Notoginseng by Multi-Sensor Information Fusion Strategy of Infrared Spectra Combined with Random Forest. *Spectrochim Acta A Mol Biomol Spectrosc* **2020**, *226*, 117619. <https://doi.org/10.1016/j.saa.2019.117619>.
- (126) Pei; Zuo; Zhang; Wang. Data Fusion of Fourier Transform Mid-Infrared (MIR) and Near-Infrared (NIR) Spectroscopies to Identify Geographical Origin of Wild Paris Polyphylla Var. Yunnanensis. *Molecules* **2019**, *24* (14), 2559. <https://doi.org/10.3390/molecules24142559>.

- (127) Ren, G.; Wang, Y.; Ning, J.; Zhang, Z. Highly Identification of Keemun Black Tea Rank Based on Cognitive Spectroscopy: Near Infrared Spectroscopy Combined with Feature Variable Selection. *Spectrochim Acta A Mol Biomol Spectrosc* **2020**, *230*, 118079. <https://doi.org/10.1016/j.saa.2020.118079>.
- (128) Anderson, N. T.; Walsh, K. B.; Subedi, P. P.; Hayes, C. H. Achieving Robustness across Season, Location and Cultivar for a NIRS Model for Intact Mango Fruit Dry Matter Content. *Postharvest Biol Technol* **2020**, *168*, 111202. <https://doi.org/10.1016/j.postharvbio.2020.111202>.
- (129) Kartakoullis, A.; Comaposada, J.; Cruz-Carrión, A.; Serra, X.; Gou, P. Feasibility Study of Smartphone-Based Near Infrared Spectroscopy (NIRS) for Salted Minced Meat Composition Diagnostics at Different Temperatures. *Food Chem* **2019**, *278*, 314–321. <https://doi.org/10.1016/j.foodchem.2018.11.054>.
- (130) Cunha, C. L.; Torres, A. R.; Luna, A. S. Multivariate Regression Models Obtained from Near-Infrared Spectroscopy Data for Prediction of the Physical Properties of Biodiesel and Its Blends. *Fuel* **2020**, *261*, 116344. <https://doi.org/10.1016/j.fuel.2019.116344>.
- (131) Benedet, L.; Faria, W. M.; Silva, S. H. G.; Mancini, M.; Guilherme, L. R. G.; Demattê, J. A. M.; Curi, N. Soil Subgroup Prediction via Portable X-Ray Fluorescence and Visible near-Infrared Spectroscopy. *Geoderma* **2020**, *365*, 114212. <https://doi.org/10.1016/j.geoderma.2020.114212>.
- (132) Chakraborty, S.; Li, B.; Weindorf, D. C.; Deb, S.; Acree, A.; De, P.; Panda, P. Use of Portable X-Ray Fluorescence Spectrometry for Classifying Soils from Different Land Use Land Cover Systems in India. *Geoderma* **2019**, *338*, 5–13. <https://doi.org/10.1016/j.geoderma.2018.11.043>.
- (133) Andrade, R.; Faria, W. M.; Silva, S. H. G.; Chakraborty, S.; Weindorf, D. C.; Mesquita, L. F.; Guilherme, L. R. G.; Curi, N. Prediction of Soil Fertility via Portable X-Ray Fluorescence (PXRF) Spectrometry and Soil Texture in the Brazilian Coastal Plains. *Geoderma* **2020**, *357*, 113960. <https://doi.org/10.1016/j.geoderma.2019.113960>.
- (134) Andrade, R.; Silva, S. H. G.; Weindorf, D. C.; Chakraborty, S.; Faria, W. M.; Mesquita, L. F.; Guilherme, L. R. G.; Curi, N. Assessing Models for Prediction of Some Soil Chemical Properties from Portable X-Ray Fluorescence (PXRF) Spectrometry Data in Brazilian Coastal Plains. *Geoderma* **2020**, *357*, 113957. <https://doi.org/10.1016/j.geoderma.2019.113957>.
- (135) dos Santos Teixeira, A. F.; Henrique Procópio Pelegrino, M.; Missina Faria, W.; Henrique Godinho Silva, S.; Gabriela Marcolino Gonçalves, M.; Weimar Acerbi Júnior, F.; Rezende Gomide, L.; Linares Pádua Júnior, A.; de Souza, I. A.; Chakraborty, S.; Weindorf, D. C.; Roberto Guimarães Guilherme, L.; Curi, N. Tropical Soil PH and Sorption Complex Prediction via Portable X-Ray Fluorescence Spectrometry. *Geoderma* **2020**, *361*, 114132. <https://doi.org/10.1016/j.geoderma.2019.114132>.
- (136) Amjad, A.; Ullah, R.; Khan, S.; Bilal, M.; Khan, A. Raman Spectroscopy Based Analysis of Milk Using Random Forest Classification. *Vib Spectrosc* **2018**, *99*, 124–129. <https://doi.org/10.1016/j.vibspec.2018.09.003>.

- (137) Liang, B.; Gao, Y.; Xu, J.; Song, Y.; Xuan, L.; Shi, T.; Wang, N.; Hou, Z.; Zhao, Y.-L.; Huang, W. E.; Chen, Z.-J. Raman Profiling of Embryo Culture Medium to Identify Aneuploid and Euploid Embryos. *Fertil Steril* **2019**, *111* (4), 753-762.e1. <https://doi.org/10.1016/j.fertnstert.2018.11.036>.
- (138) Vignoli, A.; Tenori, L.; Giusti, B.; Takis, P. G.; Valente, S.; Carrabba, N.; Balzi, D.; Barchielli, A.; Marchionni, N.; Gensini, G. F.; Marcucci, R.; Luchinat, C.; Gori, A. M. NMR-Based Metabolomics Identifies Patients at High Risk of Death within Two Years after Acute Myocardial Infarction in the AMI-Florence II Cohort. *BMC Med* **2019**, *17* (1), 3. <https://doi.org/10.1186/s12916-018-1240-2>.
- (139) Bermudez-Lopez, M.; Forne, C.; Amigo, N.; Bozic, M.; Arroyo, D.; Bretones, T.; Alonso, N.; Cambray, S.; del Pino, M. D.; Mauricio, D.; Gorriz, J. L.; Fernandez, E.; Valdivielso, J. M. An In-Depth Analysis Shows a Hidden Atherogenic Lipoprotein Profile in Non-Diabetic Chronic Kidney Disease Patients. *Expert Opin Ther Targets* **2019**, *23* (7), 619–630. <https://doi.org/10.1080/14728222.2019.1620206>.
- (140) Mascellani, A.; Hoca, G.; Babisz, M.; Krska, P.; Kloucek, P.; Havlik, J. ¹H NMR Chemometric Models for Classification of Czech Wine Type and Variety. *Food Chem* **2021**, *339*, 127852. <https://doi.org/10.1016/j.foodchem.2020.127852>.
- (141) Qian, K.; Wang, Y.; Hua, L.; Chen, A.; Zhang, Y. New Method of Lung Cancer Detection by Saliva Test Using Surface-Enhanced Raman Spectroscopy. *Thorac Cancer* **2018**, *9* (11), 1556–1561. <https://doi.org/10.1111/1759-7714.12837>.
- (142) Živanović, V.; Seifert, S.; Drescher, D.; Schrade, P.; Werner, S.; Guttman, P.; Szekeres, G. P.; Bachmann, S.; Schneider, G.; Arenz, C.; Kneipp, J. Optical Nanosensing of Lipid Accumulation Due to Enzyme Inhibition in Live Cells. *ACS Nano* **2019**, *13* (8), 9363–9375. <https://doi.org/10.1021/acsnano.9b04001>.
- (143) Seifert, S. Application of Random Forest Based Approaches to Surface-Enhanced Raman Scattering Data. *Sci Rep* **2020**, *10* (1), 5436. <https://doi.org/10.1038/s41598-020-62338-8>.
- (144) Hu, W.; Ye, S.; Zhang, Y.; Li, T.; Zhang, G.; Luo, Y.; Mukamel, S.; Jiang, J. Machine Learning Protocol for Surface-Enhanced Raman Spectroscopy. *J Phys Chem Lett* **2019**, *10* (20), 6026–6031. <https://doi.org/10.1021/acs.jpcllett.9b02517>.
- (145) Désir, C.; Bernard, S.; Petitjean, C.; Heutte, L. A Random Forest Based Approach for One Class Classification in Medical Imaging. In: Wang, F., Shen, D., Yan, P., Suzuki, K. (Eds) *Machine Learning in Medical Imaging*; 2012; pp 1–31. https://doi.org/10.1007/978-3-642-35428-1_31.
- (146) Désir, C.; Bernard, S.; Petitjean, C.; Heutte, L. A New Random Forest Method for One-Class Classification. *Lecture Notes in Computer Science* **2012**, *7626 LNCS*, 282–300.
- (147) Désir, C.; Bernard, S.; Petitjean, C.; Heutte, L. One Class Random Forests. *Pattern Recognit* **2013**, *46* (12), 3490–3506. <https://doi.org/10.1016/j.patcog.2013.05.022>.
- (148) Oliveri, P.; Malegori, C.; Casale, M. *Chemometrics and Statistics | Multivariate Classification Techniques*, Third Edit.; Elsevier, 2019; Vol. 1. <https://doi.org/10.1016/B978-0-12-409547-2.14239-8>.

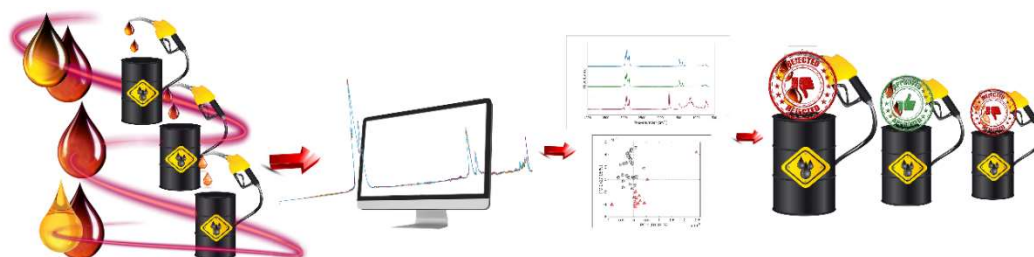
- (149) Brereton, R. G. One-Class Classifiers. *J. Chemometrics* **2011**, 25 (5), 225–246. <https://doi.org/10.1002/cem.1397>.
- (150) Oliveri, P. Class-Modelling in Food Analytical Chemistry: Development, Sampling, Optimisation and Validation Issues – A Tutorial. *Analytica Chimica Acta*. Elsevier B.V. August 22, 2017, pp 9–19. <https://doi.org/10.1016/j.aca.2017.05.013>.
- (151) Breiman, L. *Wald Lecture II Looking inside the Black Box*; 2002. <https://www.stat.berkeley.edu/users/breiman/wald2002-2.pdf> (accessed 2022-01-15).
- (152) Chen, C.; Liaw, A. *Using Random Forest to Learn Imbalanced Data*; 2004. <https://statistics.berkeley.edu/sites/default/files/tech-reports/666.pdf> (accessed 2022-01-15).
- (153) Strobl, C.; Boulesteix, A. L.; Zeileis, A.; Hothorn, T. Bias in Random Forest Variable Importance Measures: Illustrations, Sources and a Solution. *BMC Bioinformatics* **2007**, 8. <https://doi.org/10.1186/1471-2105-8-25>.
- (154) Livingston, F. Implementation of Breiman's Random Forest Machine Learning Algorithm. *Machine Learning Journal Paper*. **2005**, ECE591Q, 1–13.
- (155) Ahmadi Azqhandi, M. H.; Ghaedi, M.; Yousefi, F.; Jamshidi, M. Application of Random Forest, Radial Basis Function Neural Networks and Central Composite Design for Modeling and/or Optimization of the Ultrasonic Assisted Adsorption of Brilliant Green on ZnS-NP-AC. *J Colloid Interface Sci* **2017**, 505, 278–292. <https://doi.org/10.1016/j.jcis.2017.05.098>.
- (156) Khoshgoftaar, T. M.; Golawala, M.; van Hulse, J. An Empirical Study of Learning from Imbalanced Data Using Random Forest. In *Proceedings - International Conference on Tools with Artificial Intelligence, ICTAI*; 2007; Vol. 2, pp 310–317. <https://doi.org/10.1109/ICTAI.2007.46>.
- (157) Granitto, P. M.; Furlanello, C.; Biasioli, F.; Gasperi, F. Recursive Feature Elimination with Random Forest for PTR-MS Analysis of Agroindustrial Products. *Chemometrics and Intelligent Laboratory Systems* **2006**, 83 (2), 83–90. <https://doi.org/10.1016/j.chemolab.2006.01.007>.
- (158) Nembrini, S. On the Behaviour of Permutation-Based Variable Importance Measures in Random Forest Clustering. *J Chemom* **2019**, 33 (8). <https://doi.org/10.1002/cem.3135>.
- (159) Shi, T.; Horvath, S. Unsupervised Learning with Random Forest Predictors. *Journal of Computational and Graphical Statistics* **2006**, 15 (1), 118–138. <https://doi.org/10.1198/106186006X94072>.
- (160) Chen, S.; Plouffe, A.; Hattori, K. A Multivariate Statistical Approach Identifying the Areas Underlain by Potential Porphyry-Style Cu Mineralization, South-Central British Columbia, Canada. *J Geochem Explor* **2019**, 202, 13–26. <https://doi.org/10.1016/j.gexplo.2019.03.016>.
- (161) Poulakis, K.; Pereira, J. B.; Mecocci, P.; Vellas, B.; Tsolaki, M.; Kłoszewska, I.; Soininen, H.; Lovestone, S.; Simmons, A.; Wahlund, L. O.; Westman, E. Heterogeneous Patterns of Brain Atrophy in Alzheimer's Disease. *Neurobiol Aging* **2018**, 65, 98–108. <https://doi.org/10.1016/j.neurobiolaging.2018.01.009>.

- (162) Baron, D.; Poznanski, D. The Weirdest SDSS Galaxies: Results from an Outlier Detection Algorithm. *Mon Not R Astron Soc* **2017**, *465* (4), 4530–4555. <https://doi.org/10.1093/mnras/stw3021>.
- (163) Reis, I.; Poznanski, D.; Baron, D.; Zasowski, G.; Shahaf, S. Detecting Outliers and Learning Complex Structures with Large Spectroscopic Surveys - A Case Study with APOGEE Stars. *Mon Not R Astron Soc* **2018**, *476* (2), 2117–2136. <https://doi.org/10.1093/mnras/sty348>.
- (164) Peerbhay, K.; Mutanga, O.; Lottering, R.; Ismail, R. Mapping Solanum Mauritianum Plant Invasions Using WorldView-2 Imagery and Unsupervised Random Forests. *Remote Sens Environ* **2016**, *182*, 39–48. <https://doi.org/10.1016/j.rse.2016.04.025>.
- (165) Chaber, R.; Arthur, C. J.; Łach, K.; Raciborska, A.; Michalak, E.; Bilska, K.; Drabko, K.; Depciuch, J.; Kaznowska, E.; Cebulski, J. Predicting Ewing Sarcoma Treatment Outcome Using Infrared Spectroscopy and Machine Learning. *Molecules* **2019**, *24* (6). <https://doi.org/10.3390/molecules24061075>.
- (166) Rokach, L. Ensemble-Based Classifiers. *Artif Intell Rev* **2010**, *33* (1–2), 1–39. <https://doi.org/10.1007/s10462-009-9124-7>.
- (167) Li, H. D.; Liang, Y. Z.; Xu, Q. S.; Cao, D. S. Model Population Analysis for Variable Selection. *J Chemom* **2010**, *24* (7–8), 418–423. <https://doi.org/10.1002/cem.1300>.
- (168) Deng, B. C.; Yun, Y. H.; Liang, Y. Z. Model Population Analysis in Chemometrics. *Chemometrics and Intelligent Laboratory Systems* **2015**, *149*, 166–176. <https://doi.org/10.1016/j.chemolab.2015.08.018>.
- (169) Li, H.-D.; Liang, Y.-Z.; Xu, Q.-S. Model Population Analysis for Statistical Model Comparison. In *Chemometrics in Practical Applications*; Varmuza, K., Ed.; 2012; Vol. 23, pp 1–340. <https://doi.org/10.5772/1150>.
- (170) Afanador, N. L.; Smolinska, A.; Tran, T. N.; Blanchet, L. Unsupervised Random Forest: A Tutorial with Case Studies. *J. Chemometrics* **2016**, *30* (5), 232–241. <https://doi.org/10.1002/cem.2790>.
- (171) Zuur, A. F.; Ieno, E. N.; Smith, G. M. Analysing Ecological Data. Statistics for Biology and Health. In *Analysing Ecological Data. Statistics for Biology and Health*; Gail, M., Krickeberg, K., Samet, J., Tsiatis, A., Wong, W., Eds.; Springer: Nova York, 2007; pp 259–260. <https://doi.org/10.1007/978-0-387-45972-1>.
- (172) Brereton, R. G. Contingency Tables, Confusion Matrices, Classifiers and Quality of Prediction. *J. Chemometrics* **2021**, e3331, 1–5. <https://doi.org/10.1002/cem.3331>.
- (173) Brereton, R. G. Alpha, Beta, Type 1 and 2 Errors, Ergon Pearson and Jerzy Neyman. *Journal of Chemometrics*. John Wiley and Sons Ltd March 1, 2021. <https://doi.org/10.1002/cem.3240>.
- (174) Brereton, R. G. False Discovery Rates, Power and Related Concepts. *Journal of Chemometrics*. John Wiley and Sons Ltd June 1, 2021. <https://doi.org/10.1002/cem.3241>.
- (175) Jurman, G.; Riccadonna, S.; Furlanello, C. A Comparison of MCC and CEN Error Measures in Multi-Class Prediction. *PLoS One* **2012**, *7* (8), 1–8. <https://doi.org/10.1371/journal.pone.0041882>.

- (176) Boughorbel, S.; Jarray, F.; El-Anbari, M. Optimal Classifier for Imbalanced Data Using Matthews Correlation Coefficient Metric. *PLoS One* **2017**, *12* (6), 1–17.
<https://doi.org/10.1371/journal.pone.0177678>.

CAPÍTULO 3

O USO DE FLORESTA ALEATÓRIA NÃO SUPERVISIONADA NA DISCRIMINAÇÃO DO ÓLEO DIESEL: UM ESTUDO DE CASO DE ADULTERAÇÃO DE COMBUSTÍVEL



- Neste capítulo é apresentada uma aplicação alternativa da floresta aleatória não supervisionada para triagem de amostras em um estudo de caso de suspeita de adulteração de combustível.
- Disponibilizamos neste capítulo um algoritmo adaptado para aplicação de floresta aleatória não supervisionada de forma simples.
- O método de aprendizado de máquina pode identificar a adulteração de combustível para um número reduzido de amostras reais não rotuladas.

3. O USO DE FLORESTA ALEATÓRIA NÃO SUPERVISIONADA NA DISCRIMINAÇÃO DO ÓLEO DIESEL: UM ESTUDO DE CASO DE ADULTERAÇÃO DE COMBUSTÍVEL

Resumo

Algumas metodologias baseadas em propriedades físico-químicas e técnicas espectroscópicas têm sido utilizadas para detectar adulteração de combustível diesel. Neste trabalho, descrevemos uma aplicação alternativa do método de floresta aleatória não supervisionada (URF, do inglês *unsupervised random forest*) na análise de um caso real de fraude para um pequeno número de amostras (48 amostras) de combustível diesel de dois tipos (marítima e metropolitana). Para isso, foram avaliados parâmetros de qualidade do diesel baseados nas normas brasileiras, além de aquisição de espectros na região do infravermelho médio (FT-MIR). Identificamos nas amostras reais, possíveis amostras suspeitas via URF por meio de um modelo com MIR na região de 4000 a 650 cm^{-1} . Em contrapartida, análise de componentes principais (PCA) e modelagem independente flexível de analogia de classes (SIMCA) foram aplicadas com seleção de variáveis (iPCA), gerando quatro modelos distintos. As regiões espectrais mais relevantes foram as bandas de absorção dos grupos carbonila e aromáticos. Dezenove das amostras reais avaliadas não estavam em conformidade com a legislação brasileira (39% das amostras). Foi possível identificar as mesmas amostras por meio de análises físico-químicas e espectroscopia. O óleo de soja também foi identificado como um adulterante, mas não o único devido a variações nas bandas de absorção de grupos aromáticos. Embora os métodos de teste padrão não tenham apresentado alterações, o modelo URF ou o PCA e SIMCA nas regiões de carbonila e aromáticos podem fornecer informações de suspeita de fraude entre as amostras. A URF pode detectar a relação linear e não linear a partir do conjunto de dados, mesmo com número desbalanceado de amostras. Este método pode fornecer uma triagem eficiente para amostras não rotuladas. Ademais, foi possível uma resposta a partir de uma triagem utilizando um modelo com os espectros inteiros, enquanto o mesmo resultado exigiu a seleção de diferentes regiões espectrais ao se utilizar outros métodos como PCA e SIMCA.

Palavras-chave: diesel; adulteração; controle de qualidade; FTIR; URF; aprendizado de máquina.

3.1. Introdução

Técnicas de espectroscopia têm sido utilizadas na identificação e quantificação de adulterantes nos problemas de adulteração de combustível. Mas as semelhanças químicas entre adulterantes e combustíveis ainda podem dificultar esse processo.¹⁻⁴ Bandas de absorção nas regiões espectroscópicas do infravermelho médio (MIR) e infravermelho próximo (NIR) relacionadas a grupos funcionais e associação com análise de dados multivariados fornecem resultados robustos.⁴⁻¹³ Nesse contexto, Bassbasi *et al.*⁸ utilizaram quimiometria associada a dados MIR para investigar a adulteração de óleo de alta qualidade com óleo de baixa qualidade. Para isso, foram utilizadas três classes de óleo de motor e misturas binárias dessas amostras. Os autores aplicaram a análise de componentes principais (PCA), que forneceu uma distinção prévia entre amostras adulteradas e não adulteradas. Eles também usaram a análise discriminante de mínimos quadrados parciais (PLS-DA) na qual as misturas binárias (0 - 20 em peso%) foram preditas com 100% de precisão. PCA e PLS-DA também foram utilizados por Moura *et al.*⁴ com dados MIR e NIR para identificar adulteração de amostras de diesel com querosene. A PCA discriminou amostras de diesel em grupos com diferentes tempos de armazenamento, e com o PLS-DA foi possível classificar amostras adulteradas com 100% de precisão. Câmara *et al.*¹² utilizaram dados MIR e NIR associados ao método de resolução de curva multivariada de mínimos quadrados alternativos (MCR-ALS) para identificar a adulteração residual de óleo de fritura vegetal em diesel comercial S10 com biodiesel (8% v/v) (S10B8). Aleme *et al.*¹⁴ aplicaram PCA e análise discriminante linear (LDA) para classificar diesel puro (B0) e B4 (combustível diesel com 4% em peso de biodiesel). No entanto, este estudo destacou que os parâmetros físico-químicos e as curvas de destilação não conseguiram distinguir as amostras de acordo com o teor de biodiesel (B0 e B4), devido às semelhanças físico-químicas entre diesel e biodiesel.

A aprendizagem de máquina parece ser uma alternativa promissora para superar problemas na detecção de adulteração. Agrega precisão e robustez, mesmo com um número desbalanceado de amostras e variáveis.¹⁵⁻¹⁷ Neste contexto, Le *et al.*¹⁷ avaliaram um método de aprendizagem (*deep learning*) para selecionar as características de amostras de diesel na região do NIR. Eles combinaram dois

métodos de aprendizagem de máquina para criar um modelo para teor de aromáticos totais de 395 combustíveis diesel (dados NIR).

Dentre os métodos de aprendizado de máquina, a floresta aleatória (RF) foi recentemente aplicada a estudos de classificação de petróleo bruto,^{18,19} classificação de diesel¹⁵ e para avaliar a qualidade de biodiesel. RF é um método do tipo *ensemble*, desenvolvido por Breiman²⁰ por meio de combinação de árvores de classificação e regressão (CART) com algoritmos de *bootstrapping (bagging)*.^{20,21} Lovatti et al.¹⁸ aplicaram RF para distinguir 105 óleos brutos com base no ponto de fluidez usando um conjunto de dados de ressonância magnética nuclear (RMN). Para isso, avaliaram as amostras com ponto de fluidez máximo e propriedades físico-químicas para atribuir duas classes com base em um limite de valor ($\leq -9,0$ °C e $> -9,0$ °C) e desenvolveram modelos de RF com diferentes pré-processamento dos dados. Ademais, propuseram um método de identificação de variáveis mais importantes por meio de RF, sendo identificadas as relações dessas variáveis com a composição do petróleo e as propriedades físico-químicas.

Ranzan et al.¹⁵ compararam os desempenhos PLS-DA e RF na classificação de diesel (S10 e S100) para controle de qualidade com base no teor de enxofre. Um total de 51 amostras de diesel S100 e 12 diesel S10 foram analisadas com espectroscopia de fluorescência 2D. O conjunto de dados foi avaliado pelos modelos PCA, PLS-DA e RF para análise qualitativa com seleção de variáveis. O PLS-DA obteve uma boa classificação com duas variáveis latentes que explicam quase 99% da variância dos dados de 27 variáveis originais, com uma boa classe preditiva no conjunto de testes. Para o modelo de RF, todos os dados foram utilizados como entrada, sem seleção de variáveis, e todos os subconjuntos de teste foram classificados corretamente. No entanto, os autores não puderam confirmar uma correlação direta entre as variáveis selecionadas e compostos de enxofre ou outra substância removida durante a hidrodessulfurização.

Normalmente, o aprendizado de máquina supervisionado requer conjuntos de dados rotulados para treinar o modelo, o que pode não ser prático em um ambiente real. Dessa forma, a floresta aleatória não supervisionada tem sido sugerida como um método alternativo.²²⁻²⁷ URF é um método adicional de aprendizado de máquina para reconhecimento de padrões nos dados. Este método usa a floresta aleatória como uma análise não supervisionada com base na presença de uma estrutura nos

dados. Essa estrutura permite a distinção de uma versão aleatória do conjunto de dados original. Essa abordagem tem sido empregada em estudos biológicos para identificar biomarcadores preditivos de câncer ou para extrair características para reconhecimento de padrões.^{24,28} URF também tem sido aplicada em estudos geoquímicos,²³ sensoriamento do meio ambiente²⁷ e astronomia.^{25,26} No entanto, este método não supervisionado ainda é pouco explorado em química analítica.²²

Neste estudo, propomos o uso da URF como um método de aprendizagem estatística associado à espectroscopia para triagem de amostras reais de diesel, não rotuladas, as quais fomos solicitados a analisar. As amostras necessitaram de análise com base nas resoluções da Agência Nacional do Petróleo, Gás Natural e Biocombustíveis (ANP) para o diesel metropolitano (DM) e o diesel marítimo A (DMA).^{29,30} Foram avaliados parâmetros físico-químicos por meio de métodos padronizados e realizados modelos de classificação a partir da espectroscopia MIR. Além disso, os mesmos parâmetros foram avaliados em amostra de diesel puro e em amostras adulteradas com solvente orgânico terebentina e óleo de soja não refinado para associar o comportamento dessas amostras com as do estudo de caso.

3.2. Metodologia

Neste estudo, 48 amostras de diesel (36 DMA e 12 DM S500) sem biodiesel foram recebidas pelo Laboratório de Pesquisa e Desenvolvimento de Metodologias para Análise de Petróleo (LabPetro), da Universidade Federal de Espírito Santo (UFES), para identificar não conformidades. Utilizou-se também uma amostra de diesel sem biodiesel como referência interna. Este diesel foi usado para preparar mistura binária com óleo de soja não refinado e solvente terebentina. Para isso, o diesel padrão (DP) foi misturado com terebentina (T) e óleo de soja não refinado (S), nas concentrações de 1, 5, 10 e 20 % v/v. Todas as amostras foram analisadas em triplicata pelos métodos aprovados pela *American Society for Testing and Materials* (ASTM), seguindo as diretrizes da ANP.

Em seguida, os espectros MIR foram adquiridos para se avaliar os perfis espectrais de todas as amostras, gerando-se os conjuntos de dados: 1º) dados espectrais das amostras de diesel para o estudo de caso, bem como os resultados de seus parâmetros; 2º) dados espectrais das misturas binárias de PD com óleo de soja não refinado e terebentina. Os espectros foram pré-processados e avaliados

pelos métodos de reconhecimento de padrões não supervisionados PCA e URF. As variáveis foram selecionadas por meio da análise de componentes principais por intervalos (iPCA). A região MIR identificada pelo iPCA foi utilizada para o modelo de classe única modelagem independente e flexível de analogias de classe (SIMCA) para classificar as amostras de acordo com a classe-alvo: amostras conformes.

3.2.1. Caracterização dos óleos

Os métodos de ensaio foram conduzidos de acordo com as normas ASTM: densidade D4052³¹, ponto de fulgor D93³², número de acidez total (NAT) D664³³, teor total de enxofre D4294³⁴, viscosidade cinemática D7042³⁵ e ponto de fluidez D97³⁶.

3.2.1.1 Densidade

A densidade do óleo foi determinada de acordo com a norma ASTM D4052³¹. Utilizou-se um viscosímetro digital Anton Paar (modelo Stabinger SVM 3000) com limite de detecção de 0,0002 g cm⁻³ a 20 °C. Aproximadamente 5 mL da amostra foram injetados na célula de medição do viscosímetro.

3.2.1.2 Ponto de fulgor

O ponto de fulgor foi determinado de acordo com a norma ASTM D93³². Um volume de 70 mL foi transferido para o analisador de ponto de fulgor, ISL modelo FP93. O copo de teste foi aquecido, e, em seguida, agitado em intervalos (5 °C, 1 min e 90-120 rpm, respectivamente). Uma fonte de ignição foi aplicada ao copo de ensaio em intervalos regulares de 1 °C, começando de 23 a 28 °C abaixo do ponto de fulgor previsto. A agitação no sistema só foi interrompida após a detecção de um *flash* distinto. O ponto de fulgor foi relatado como a temperatura mais baixa à pressão barométrica de 101,3 kPa. Neste ensaio, a fonte de ignição provoca a combustão dos vapores da amostra e os resultados são expressos em °C.

3.2.1.3 Número de acidez total

NAT foi realizado por titulação potenciométrica de acordo com a norma ASTM D664³³. 20,0 g ($\pm 0,5$) da amostra de óleo foi dissolvido em solução com tolueno (50%), isopropanol (49,5%) e água ultrapura (0,5%). Em seguida, a titulação foi

realizada com uma solução alcoólica de hidróxido de potássio (KOH) em titulador automático, fabricado pela Metrohm (modelo 836 Titrando).

3.2.1.4. Teor total de enxofre

O teor total de enxofre no óleo foi determinado de acordo com a norma ASTM D4294³⁴, utilizando-se espectrometria de fluorescência de raios X por dispersão de energia em um analisador automático fabricado pela HORIBA (modelo SFLA-2800). Três curvas de calibração foram construídas utilizando padrões para enxofre em óleo mineral com concentrações de 0,005-0,100 (wt%), 0,05-1,00 (wt%) e 0,3-4,0 (wt%). Após a leitura da amostra, seleciona-se automaticamente a curva de calibração de acordo com a amostra analisada, determinando assim o teor total de enxofre do óleo.

3.2.1.5 Viscosidade cinemática a 60 °C

A viscosidade cinemática foi determinada pela injeção de uma amostra em um viscosímetro digital, Stabinger (modelo SVM 3000), fabricado pela Anton Paar. Este parâmetro foi medido a 60 °C e seguindo os padrões ASTM D7042³⁵.

3.2.1.6 Ponto de fluidez

O ponto de fluidez foi determinado de acordo com a norma ASTM D97³⁶. Foi realizado em um modelo automático (CPP 5GS) pelo fabricante *Grabner Instruments*. A análise foi realizada entre -57 e 51 °C. A amostra foi inserida em um tubo de vidro dentro da câmara de resfriamento. A cada intervalo de 3,0 °C, o fluxo da amostra foi verificado.

3.2.2. Espectroscopia MIR

Os espectros de todas as amostras de diesel e as misturas PD:S e PD:T foram obtidos em triplicata, com 64 varreduras e resolução de 4 cm⁻¹ em espectrômetro Perkin Elmer (Spectrum 400), utilizando o modo de absorbância e acessório reflectância total atenuada (ATR, Zn-Se). Para cada análise, foram utilizados aproximadamente 0,5 mL da amostra sem preparo prévio com tempo médio de 3 min para cada aquisição.

Os espectros obtidos compõem dois conjuntos de dados com a média das triplicatas. O primeiro conjunto compreende as 48 amostras de diesel do caso de

estudo ($\mathbf{X}_{(48,1676)}$); o segundo, dados espectrais de DP misturado com S e T ($\mathbf{X}_{(11,1676)}$).

3.2.3 Análise multivariada dos dados

3.2.3.1 Pré-processamento e modelo PCA

Os espectros foram pré-processados para correção de linha de base e suavização Savitzky-Golay³⁷ (janela 5 pontos, polinômio 2º grau). Além disso, alguns pré-processamentos foram avaliados, como normalização de dados, variação normal padrão (SNV)³⁸, primeira e segunda derivadas (janela de 9 pontos, polinômio de 2º grau) e correção de dispersão multiplicativa (MSC).³⁹ Essa etapa de pré-processamento é importante para remover variações nos espectros que não se referem às características da amostra. Todo o processamento e modelos foram desenvolvidos utilizando o software Matlab 2013® com funções como o iToolbox⁴⁰ e scripts próprios para modelos PCA^{41,42} e RF.^{18,19,22}

A PCA é um dos métodos mais aplicados em quimiometria, utilizado para reduzir a dimensão dos dados e determinar padrões amostrais. PCA transforma linearmente o conjunto original de variáveis em um conjunto menor de variáveis não correlacionadas que contém a maior parte das informações do conjunto original.^{41,42} Os espectros MIR fornecem mais de 1000 variáveis e uma redução é importante para extrair informações relevantes e distinguir amostras. Ademais, iPCA⁴⁰ foi aplicada como método de seleção de variáveis para a análise não supervisionada. Esta técnica divide conjuntos de dados em vários intervalos (variáveis), calcula modelos de PCA para cada intervalo e apresenta resultados em gráficos de *scores*. Assim, é possível interpretar o conjunto de amostras de acordo com diferentes regiões espectrais. Para isso, o espectro foi dividido em 20 intervalos equidistantes e um modelo PCA gerado para cada intervalo foi avaliado. Após a seleção da região espectral, um novo modelo PCA foi construído utilizando variáveis dentro da faixa selecionada.

3.2.3.2 Modelo URF

URF foi utilizada como um método alternativo de reconhecimento de padrões para triagem das amostras neste estudo de caso. A amostragem pelo *bootstrap*,

realizada neste método, oferece diferentes possibilidades para conjunto de treinamento na formação das árvores de classificação. Isso é importante em análises espectrais de diesel ou outros derivados de petróleo, uma vez que geralmente resultam em número de variáveis superiores ao número de amostras.^{18,19,43} Neste trabalho, utilizamos URF a partir de um algoritmo próprio, adaptado de acordo com metodologia de Shi e Horvatgh⁴⁴ e Afanador et al.²² Este algoritmo tem uma adaptação para avaliar internamente a taxa de erro do conjunto OOB. A partir dessa adaptação, o algoritmo somente segue com a análise pela matriz de proximidade e projeção das amostras pela análise de coordenadas principais, caso seja atendido o critério estabelecido de 70% de exatidão. Esta exatidão é obtida a partir de um modelo supervisionado gerado entre os dados originais e os *outliers* artificiais. Apesar de na literatura este valor de exatidão estar em torno de 50%,^{20,22} adotamos um valor de corte maior (70%), com base em estudos anteriores, para garantir modelos com maior relação com a resposta de interesse. A análise pela PcoA utilizou a distância Euclidiana para projeção dos resultados da URF em um espaço de menor dimensão.

3.2.3.3 Modelo supervisionado via SIMCA

Aplicamos o método SIMCA^{45,46} usando o algoritmo *classification toolbox*⁴⁷. Para isso, os conjuntos de dados com o mesmo pré-processamento e cada região espectral selecionada pela iPCA foram centrados na média e utilizados para a construção de um modelo com validação cruzada k-fold ($k = 5$). A partir dessa validação cruzada, escolhemos 2 PCs para a classe alvo, atribuída a amostras conformes. Foram alocadas no conjunto de dados de teste amostras com comportamento suspeito observado pela análise não supervisionada. Outras amostras conformes e não conformes foram divididas aleatoriamente em aproximadamente 65% de treinamento e aproximadamente 35% de teste.

3.2.3.4 Avaliação dos modelos de classificação

Os modelos de classificação foram avaliados pelas métricas calculadas a partir da matriz de confusão (Tabela 3.1).⁴⁸ O modelo SIMCA foi avaliado por sensibilidade (*Sens*), especificidade (*Espec*), exatidão (*Exat*), precisão (*Prec*) e taxa de erro (ER, do inglês, *error rate*). Sensibilidade expressa a capacidade do modelo

classificar corretamente as amostras pertencentes à classe alvo como sendo desta classe, ou seja, taxa de verdadeiro positivo (TVP). Por outro lado, a especificidade é a capacidade do modelo identificar amostras que não pertencem a esta classe e classificadas como tal, ou seja, taxa de verdadeiro negativo (TVN). A exatidão é a razão entre as amostras corretamente classificadas e número total de amostras. Já a precisão (*Prec*) é a razão entre o número de amostras pertencentes à classe-alvo corretamente atribuídas a esta classe e o número total de amostras atribuídas à classe.⁴⁷⁻⁴⁹

Tabela 3.1 Características de desempenho para avaliação de modelos de classificação

Métrica	Equação	
Sensibilidade	$Sens = \frac{VP}{VP + FN}$	(Eq. 3)
Especificidade	$Espec = \frac{VN}{FP + VN}$	(Eq. 4)
Exatidão	$Exat = \frac{VP + VN}{VP + FN + FP + VN}$	(Eq. 5)
Taxa de Falso Positivo	$TFP = \frac{FP}{FP + VN}$	(Eq. 6)
Taxa de Falso Negativo	$TFN = \frac{FN}{FN + VP}$	(Eq. 7)
Precisão	$Prec = \frac{VP}{VP + FP}$	(Eq. 8)

3.3. Resultados e Discussão

Neste estudo, avaliamos a presença de adulteração entre dados reais de diesel não rotulados, não apenas com óleo de soja como esperado. Mas uma triagem com o modelo URF destacou amostras suspeitas, que não foram identificadas pelos testes físico-químicos e análise PCA, mas apresentaram similaridade com mistura de diesel com adulterante aromático.

3.3.1. Caracterização dos óleos

Os parâmetros físico-químicos ponto de fulgor, ponto de fluidez, densidade, viscosidade cinemática (60 °C), número total de acidez e teor de enxofre foram definidos para as 48 amostras de diesel recebidas para identificá-las como conformes ou não conformes de acordo com os padrões da ANP^{29,30}. Além disso, a amostra de diesel puro DP, os adulterantes T e S e suas misturas binárias DP:T e DP:S também foram avaliadas por meio de testes de parâmetros físico-químicos.

Para o ponto de fulgor de diesel S500 metropolitano (DM) e diesel marítimo (DMA), a legislação da ANP estabelece os valores mínimos de 38 e 60 °C, respectivamente.^{29,30} O ponto de fulgor do solvente terebentina (T) varia de 38 a 52 °C, enquanto o ponto de fulgor do óleo de soja tem o máximo de 230 °C. O ponto de fulgor das amostras de DM resultou acima de 38 °C (Figura 3.1a), seguindo as especificações da ANP (38 °C). Em relação às amostras de DMA, dois mediram os valores de ponto de fulgor abaixo de 60 °C (Figura 3.1a), ou seja, o que não está em conformidade com as especificações (60 °C). As misturas DP:T apresentaram uma diminuição em seus valores de ponto de fulgor com um aumento na concentração de terebentina. Isso acontece devido ao ponto de fulgor mais baixo desse solvente. No entanto, em todos os níveis de concentração, seus valores de ponto de fulgor foram superiores a 38 °C, em conformidade com as especificações regulamentares. Por outro lado, a mistura DP:S apresentou aumento em seu ponto de fulgor, ao aumentar a concentração de óleo de soja. Este efeito é devido ao maior ponto de fulgor do óleo de soja quando comparado ao diesel.

As amostras de diesel (DMA e DM) apresentaram um ponto de fluidez (Figura 3.1b) variando de -3 a -24 °C; e -18 a 9 °C, respectivamente. Este parâmetro deve ter um valor máximo de 6 a 0 °C para DMA e DM. Assim, quatro amostras (DM) apresentaram o ponto de fluidez medido fora das especificações. As misturas apresentaram o ponto de fluidez diminuindo quando a concentração de adulterantes aumentou. A temperatura de -27 °C foi o ponto de fluidez mínimo na amostra de mistura com a concentração máxima de terebintina; -16 °C foi o ponto de precipitação mínimo na amostra com a concentração máxima de óleo de soja.^{29,30}

Da mesma forma, a densidade deve variar de 0,8015 a 0,865 gcm⁻³ a 20 °C para DM e DMA. Todas as amostras apresentaram densidade dentro da faixa de

especificação regulatória, com valores de 0,829 a $(0,866 \pm 0,002)$ gcm^{-3} a 20 °C (Figura 3.1c). T e DP) apresentaram densidade máxima de $(0,788 \pm 0,002)$ gcm^{-3} e $0,833 \pm 0,002$ gcm^{-3} a 20 °C, respectivamente. As misturas apresentaram valores de densidade de acordo com as especificações da ANP. Para viscosidade cinemática, o valor especificado é de 1,5 a 6,0 $\text{mm}^2 \text{s}^{-1}$ e 2,0 a 6,0 mm^2s^{-1} para DMA e DM, respectivamente. As amostras de diesel (Figura 3.1d), DP e misturas DP:T em todos os níveis de concentração apresentaram esse parâmetro de acordo com as especificações regulatórias. Por outro lado, as misturas DP:S apresentaram valores de viscosidade de 2,02; 5,40; 8,69 e 10,33 mm^2s^{-1} . Esse comportamento pode ser explicado pelo maior valor de viscosidade cinemática do óleo de soja (24,00 $\text{mm}^2\cdot\text{s}^{-1}$).

Para acidez, as amostras de DMA apresentaram valores acima de 0,5 mg KOH g^{-1} (Figura 3.1e), não conformes às especificações da ANP. Entre as misturas, DP:S apresentou 0,08, 0,32, 0,53 e 0,85 mg KOH g^{-1} para cada nível de concentração de óleo de soja, respectivamente. É importante destacar que o óleo de soja tem um valor de NAT de 1,20 mg KOH g^{-1} . Assim, a presença dessa substância no diesel deve aumentar o parâmetro NAT. Finalmente, para o teor total de enxofre (Figura 3.1f), apesar de uma amostra de DM apresentar uma suposta não conformidade ($0,0518 \pm 0,0035$ wt%), esse valor está dentro do limite estabelecido pelo método. O teor máximo de enxofre deve ser de 0,050% para DM e 0,500% para DMA. Outras amostras – DP, e suas misturas, DP:T e DP:S – também apresentaram teor de enxofre dentro das especificações regulatórias. Para distinguir as amostras adulteradas, foram avaliados os dados da espectroscopia juntamente com a análise multivariada dos dados. As informações sobre as propriedades físico-químicas de NAT, ponto de fulgor e teor de enxofre foram usadas como referência para categorizar as amostras em duas classes: conforme ou não conforme.

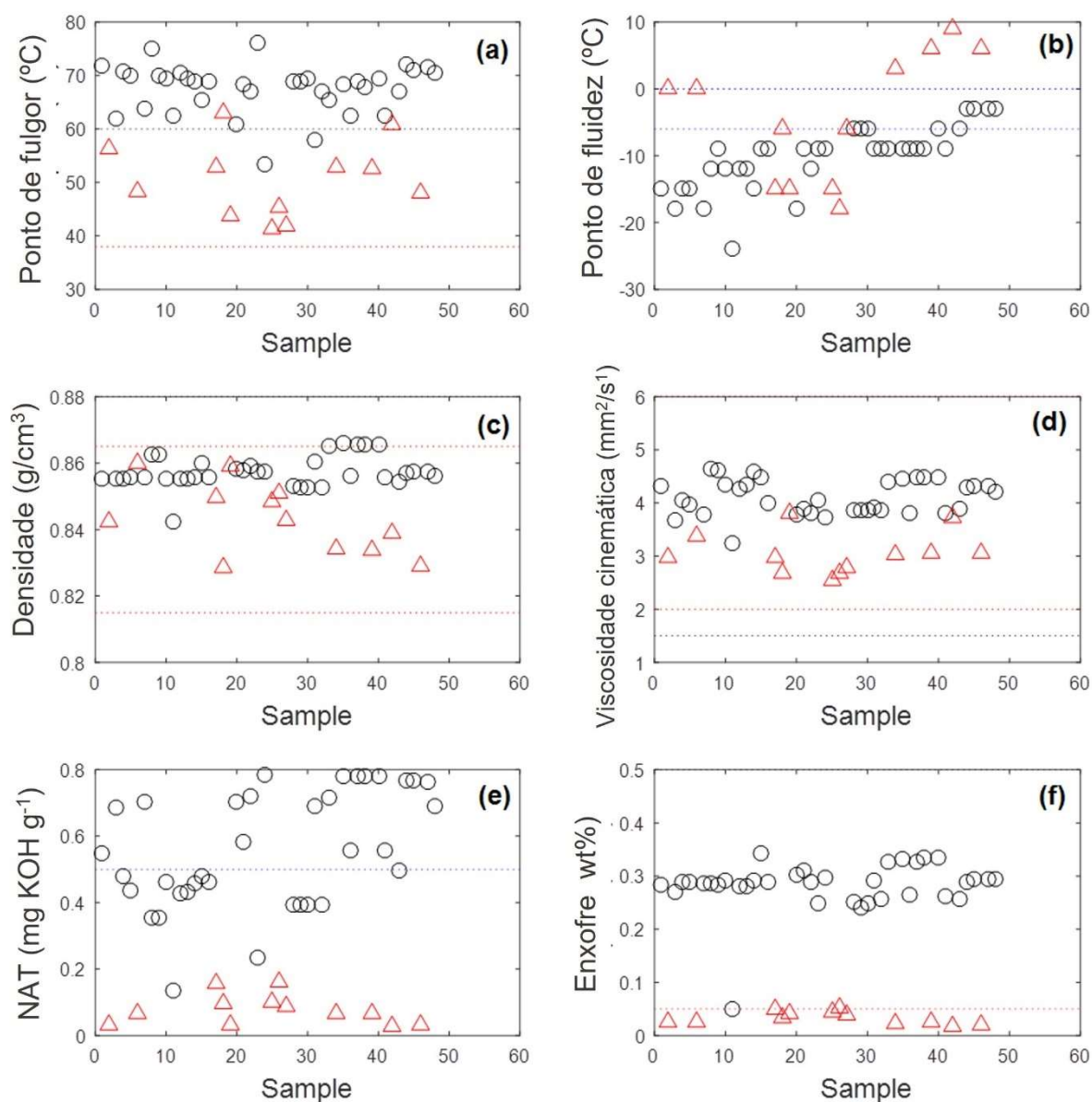


Figura 3.1 Avaliação univariada dos métodos de ensaio das propriedades físico-químicas do óleo diesel metropolitano S500 (Δ , DM) e óleo diesel marítimo A ($\circ$, DMA). (a) ponto de fulgor; (b) ponto de fluidez; (c) densidade; (d) viscosidade cinemática; (e) NAT; (f) teor de enxofre.

3.3.2 Espectroscopia MIR e análise multivariada de dados

Os espectros de diesel (Figura 3.2(A)) apresentam três regiões de absorção significativa. Bandas de alta intensidade correspondentes a vibrações de deformação axial simétricas e assimétricas da ligação C-H de metil (CH_3) e metileno (CH_2) são vistas a $3000\text{-}2840\text{ cm}^{-1}$ (Figura 3.2(B)). A $1500\text{-}1300\text{ cm}^{-1}$ (Figura 3.2(C)), as bandas de intensidade intermediária correspondem às vibrações de deformação angular da ligação C-H de etileno (C_2H_5) e metileno. Há uma menor

absorção na região entre 1000 e 650 cm^{-1} (Figura 3.2(D)).^{12,50,51} Apesar da menor intensidade, essas bandas contêm informações relevantes. A região de 1300 a 900 cm^{-1} (impressão digital) fornece uma característica padrão molecular e é única entre os perfis moleculares. No entanto, o padrão de absorção nesta região é mais complexo em geral, com bandas sobrepostas por causa das interações entre os modos vibracionais.^{50,51} Ainda na impressão digital, a absorção correspondente à vibração de deformação angular fora do plano de ligação C-H de anéis aromáticos apresenta bandas a 900-650 cm^{-1} . Além disso, o enxofre pode estar presente no diesel como mercaptano e sulfeto, com esta presença observada no espectro MIR por bandas de baixa intensidade nas regiões 2600-2550 cm^{-1} (alongamento axial da ligação S-H) e 700-650 cm^{-1} (alongamento axial da ligação C-S).⁵⁰

Além disso, bandas mais próximas de 1600 cm^{-1} correspondem a vibrações de deformação axial de ligação carbono-carbono de compostos aromáticos.^{50,51} As amostras avaliadas neste trabalho também contêm algumas bandas de intensidade intermediária a 650 cm^{-1} e de baixa intensidade a 1600 cm^{-1} , que podem corresponder à presença de compostos aromáticos.^{50,51} O combustível diesel contém hidrocarbonetos aromáticos policíclicos (PAH, do inglês, *polycyclic aromatic hydrocarbon*) com maior número de anéis fundidos em uma molécula do que outras frações de petróleo de ebulição mais baixa, como a gasolina, por exemplo. O perfil PAH do diesel é diferente do querosene e outros solventes petroquímicos.^{8,11,50,52} A adulteração do diesel com solvente aromático pode não alterar suas propriedades físico-químicas, como densidade ou viscosidade, devido à semelhança e miscibilidade. No entanto, se os solventes aromáticos tiverem um número menor ou maior de anéis fundidos, isso pode causar uma alteração nos sinais espectroscópicos por diferentes técnicas.^{50,51}

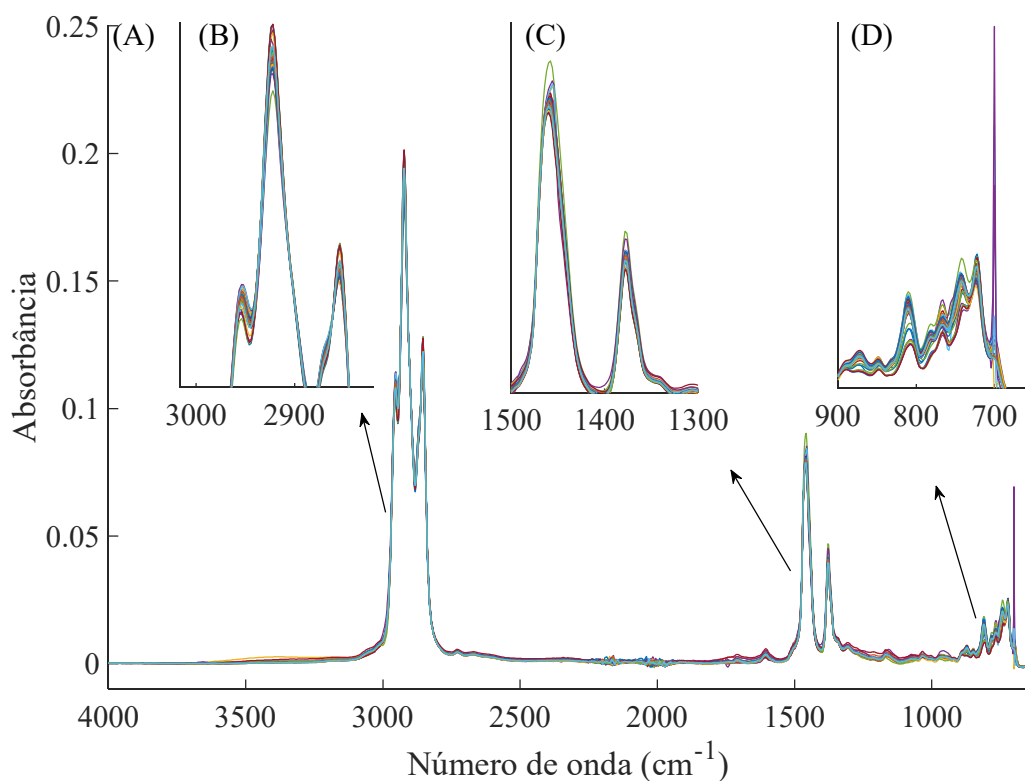


Figura 3.2 Espectros normalizados (A) de 48 amostras de diesel de estudo de caso. Bandas características do perfil espectral de diesel: 3000 - 2800 cm^{-1} (B); 1500 - 1300 cm^{-1} (C); e 900 - 650 cm^{-1} (D).

Como a maioria das amostras não conformes foi identificada pelo parâmetro NAT (19 amostras não conformes), e pela correlação deste parâmetro com a composição química, observou-se discrepância no espectro médio das amostras conformes e não conformes (Figura 3.3(A)). A primeira região discrepante entre as amostras refere-se à impressão digital (650 a 1000 cm^{-1}) (Figura 3.3(B)). Amostras com valores de NAT fora das especificações regulatórias apresentam maior absorvância nesta região, com influência dos hidrocarbonetos aromáticos, e da região de 1600 cm^{-1} .^{50,51} Essa diferença pode levantar dúvidas sobre a possibilidade de ser uma variação da linha de base espectral. No entanto, é ainda mais discrepante em torno de 1700 cm^{-1} (Figura 3.3(C)), onde a pequena faixa está presente apenas nos espectros das amostras com maior NAT.

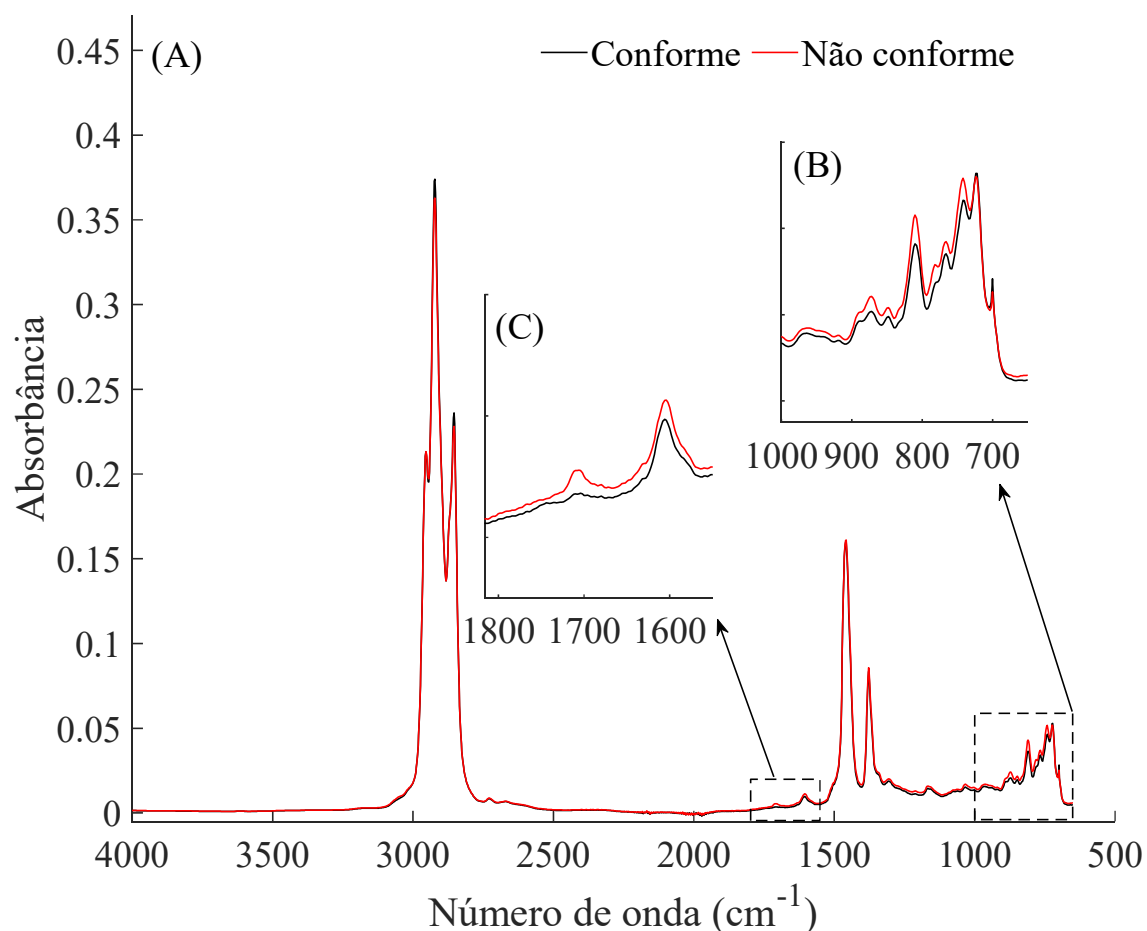


Figura 3.3 Espectro médio das amostras de diesel do caso de estudo identificadas como amostras conformes e amostras não conforme para o parâmetro TAN (ANP). (A) espectro completo; (B) região destacada da impressão digital de 1000 a 650 cm⁻¹; (C) região destacada de 1800 a 1600 cm⁻¹.

Os espectros de DP, T e S são mostrados na Figura 3.4. Os espectros das misturas DP:T e DP:S são mostrados nas Figuras S1 e S2 (Capítulo 7). Em geral, não se observam discrepâncias entre os espectros de DP (Figura 3.4(A)) e T (Figura 3.4(B)), pois possuem composição semelhante. Em contraste, os espectros de óleo de soja apresentam um perfil diferente, como esperado. As duas bandas com alta intensidade a 2855 e 2925 cm⁻¹ foram atribuídas à ligação C-H de metil (CH₃) e metileno (CH₂), associada à banda de 1380 cm⁻¹ com baixa intensidade, e 1462 cm⁻¹ (maior intensidade), ambos grupos metileno. A primeira banda a 725 cm⁻¹ de intensidade intermediária pode estar relacionada ao alceno. As bandas de 1100, 1165 e 1241 cm⁻¹ foram atribuídas à deformação axial da ligação C-O dos triacilgliceróis. Além disso, há uma faixa de maior intensidade em torno de 1700 cm⁻¹ que está ligada à vibração de alongamento das bandas de carbonila. A banda com baixa intensidade a 3011 cm⁻¹ corresponde ao trecho C-H sp². Essas

bandas são típicas do óleo de soja quando associadas a outras bandas na região de $1000\text{-}650\text{ cm}^{-1}$ (olefina, ligações C-H de deformação angular fora do plano).^{4,12,50,51}

A 1160 cm^{-1} há uma tendência linear com o aumento da concentração de óleo de soja na mistura DP:S (Figura S2(A)), além da banda de alta intensidade a 1700 cm^{-1} (Figura S1 e S2(A)). A distinção desta banda pode revelar adulteração em diesel, com óleo vegetal, ou a presença de biodiesel, uma vez que também corresponde ao modo de estiramento de C=O de éster alifático saturado. Além disso, bandas na região de $1196\text{-}1377\text{ cm}^{-1}$ também mostram informações atribuídas à presença de um éster.^{3,4,12,50,51,53}

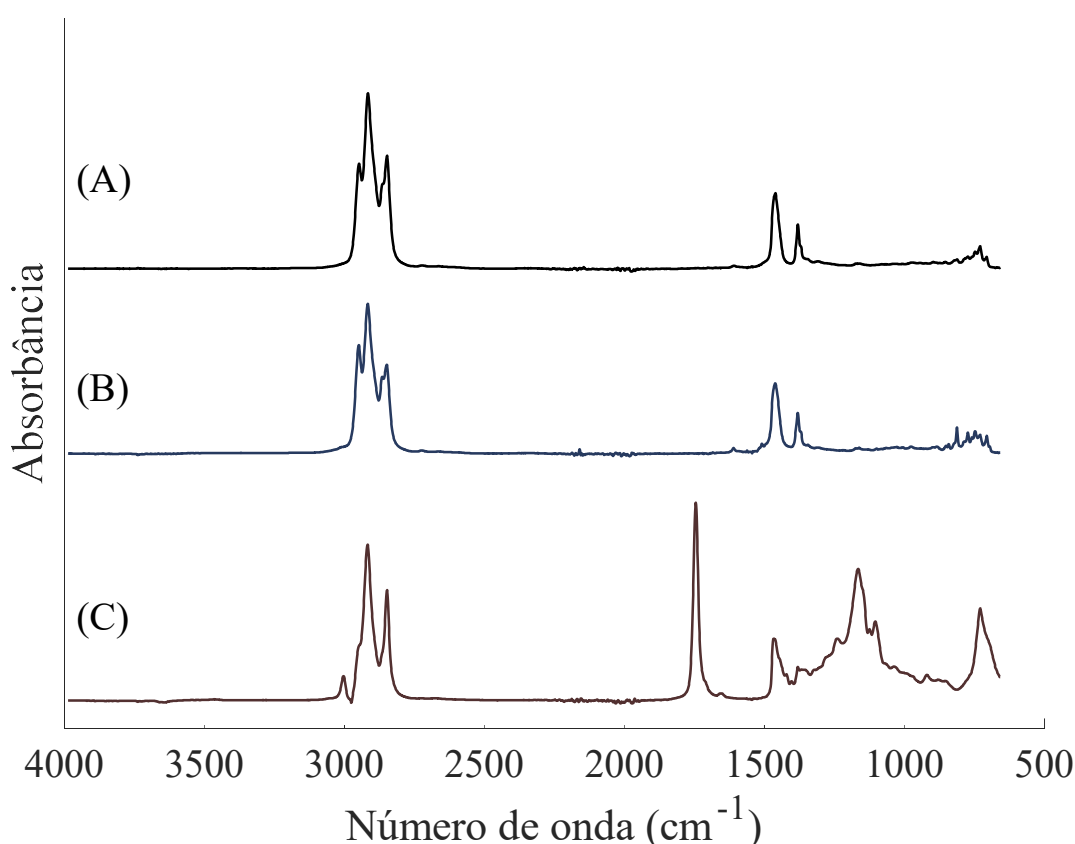


Figura 3.4 Espectros MIR na região de $650\text{ - }4000\text{ cm}^{-1}$ de (A) amostra de diesel puro adotada como referência (DP); (B) terebintina (T); (C) óleo de soja (S).

Esses dados espectrais de diesel dopado (DP:T e DP:S) foram avaliados por PCA (Figura S3), sendo que a PC1 foi suficiente para distinguir as misturas de óleo de soja e DP:S, exceto apenas a concentração de DP:S a 1% v/v com a região de $1700\text{-}1750\text{ cm}^{-1}$ (grupo carbonila) com maior *loading* (Figura S3(C)). Embora a terebintina tenha sido distinguida de outras amostras por PC2 (Figura S3), a mistura binária DP:T não foi distinguida da DP. Isso pode ser explicado pelas semelhanças

entre seus espectros e pode sugerir limitações para identificar o tipo de adulteração no combustível diesel. Ao mesmo tempo, a região da impressão digital contém a maior influência para PC2 (Figura S3(D)), principalmente estiramento de ligações =C–H de aromáticos ($900\text{--}690\text{ cm}^{-1}$).⁵⁰

Nesse sentido, o modelo PCA das 48 amostras de diesel deste estudo de caso utilizando espectros completos (1676 variáveis) obteve $92,23\%$ da variância explicada. No entanto, não foi possível observar um agrupamento entre as amostras, não sendo mostrado este resultado. No entanto, o modelo PCA desenvolvido para a região de $1816\text{--}1650\text{ cm}^{-1}$ (intervalo selecionado pela iPCA, Figura S4 e S5) foi escolhido inicialmente por distinguir as amostras em dois grupos de escores (Figura 3.5). Em um dos grupos, amostras não conformes são projetadas no quadrante positivo para PC1 (Figura 3.55(A)) sendo descritas pela região $1700\text{--}1750\text{ cm}^{-1}$ (estiramento de ligação de carbonila) (Figura 3.5 B - D). A influência dessas variáveis para distinguir a não conformidade das amostras sugere a presença de adulterantes com este grupo funcional e, conseqüentemente, uma maior acidez, o que não seria esperado para amostras de diesel deste estudo de caso. O modelo PCA nessa faixa espectral selecionada (84 variáveis) foi suficiente para identificar esse tipo de adulteração com 91% da variância explicada nas três primeiros PCs.

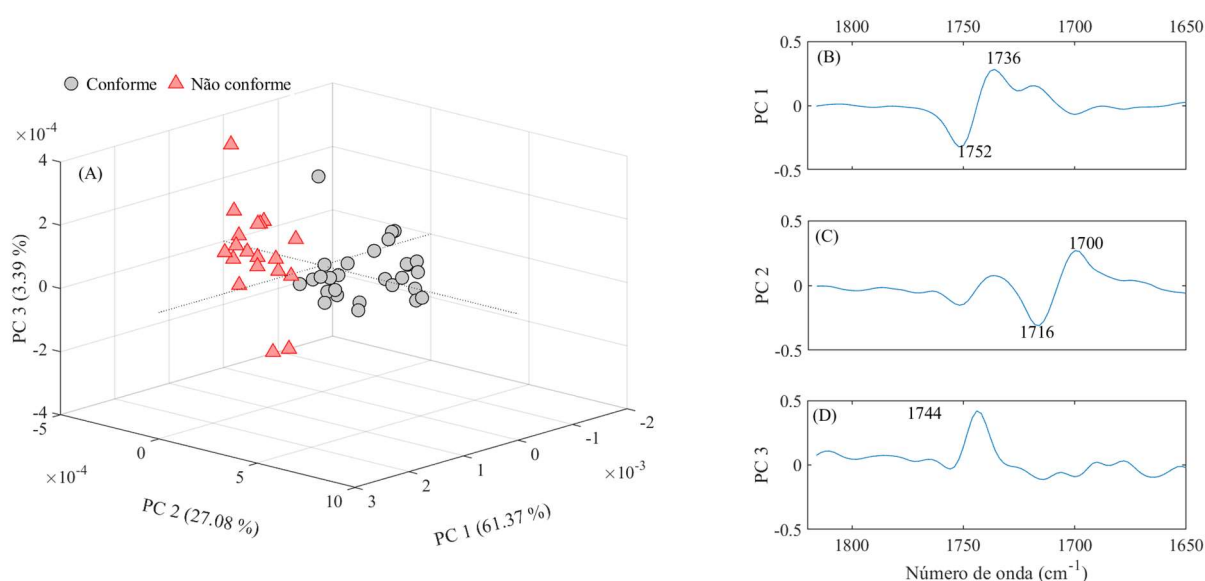


Figura 3.5 Gráficos de *scores* (A) e *loadings* (B – D) de modelo PCA (PC1 versus PC2 e PC3) de espectros MIR de 48 amostras de diesel do estudo de caso, a partir das variáveis selecionadas de 1816 a 1650 cm^{-1} .

Este resultado alinha-se com a literatura, como Corgozinho et al.⁵⁴, que utilizaram espectrofluorimetria, PCA e método supervisionado (PLS-DA) para classificar diesel e misturas com óleo de soja residual em concentrações de 0,5 a 25% em massa. Além disso, a PCA associada aos dados MIR foi utilizada por Pimentel et al.⁵⁵ para identificar o combustível diesel e misturas com biodiesel ou óleo de soja. Mas, o modelo PCA não distinguiu as misturas de diesel/éster, combustível diesel e diesel/éster/óleo, embora aplicassem a seleção de variáveis que resultou na região espectral de 1700 a 1800 cm^{-1} . No entanto, esses artigos mencionados acima avaliaram misturas em laboratório, posteriormente relacionadas a amostras comerciais. Neste estudo de caso são avaliadas amostras não rotuladas, recebidas para análise com base em suspeita de um problema real de fraude.

Outros três modelos de PCA foram construídos com intervalos designados pelas divisões de iPCA. As três regiões relacionadas aos compostos aromáticos foram escolhidas devido ao perfil das amostras em gráficos de *scores* em cada intervalo (Figura S5). As regiões selecionadas foram 1500 a 1650 cm^{-1} (Figuras S6A e S6B), 1150 a 1300 cm^{-1} (Figuras S6C e S6D) e 650 a 950 cm^{-1} (Figura S6E e S6F). As regiões selecionadas têm informações relevantes sobre compostos aromáticos devido às correlações exclusivas ou não coincidentes em 1607 cm^{-1} (axial assimétrico) e 694 cm^{-1} (simétrico angular fora de plano).^{4,12,50,52} As regiões com maiores valores de *loadings* nesses modelos foram: deformação axial de aromático (1488 cm^{-1}); C=C estiramento aromático (1600 cm^{-1}); deformação axial de C–O (1166 e 1250 cm^{-1}); e =C–H alongamento aromático (690 a 800 cm^{-1}).^{3,4,12,50,51,53}

Por outro lado, um modelo URF foi construído com toda a região espectral pré-processada com primeira derivada. Neste modelo presume-se que: se as amostras seguirem uma mesma rota até um nó terminal, em diferentes classificadores, elas são comparáveis em sua composição química. A PCoA foi utilizada para mostrar as diferenças em menor dimensão, ao invés das dimensões superiores da URF. É possível observar que há alguns grupamentos de amostras conformes, em valores de coordenadas positivas de Pco2 (Figura 3.6(A)) e coordenadas positivas de Pco1 (Figura 3.6(B)). Amostras não conformes também são apresentadas em pequenos grupos ou pares de amostras. Entre as amostras não conformes, algumas são projetadas próximas à projeção das amostras não conforme, indicando maior similaridade. Todas as amostras não conformes são

posicionadas a partir de valores de coordenadas negativas de PCo1 na Figura 3.6(B), onde também são projetadas as amostras do teste de adulteração DP:T (adulteração 1) e DP:S (adulteração 2).

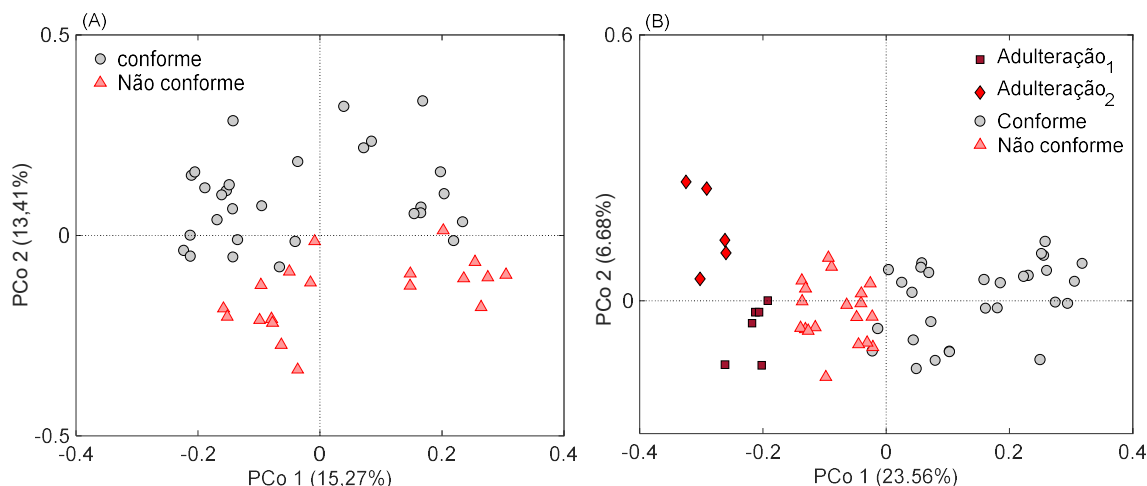


Figura 3.6 Gráfico PCoA das coordenadas (distância Euclidiana) a partir da matriz de dissimilaridade de 48 amostras de diesel do caso de estudo no modelo URF além das misturas com terebentina (Adulteração 1) e óleo de soja (Adulteração 2).

Devido ao comportamento das amostras reais nos modelos URF e PCA (regiões selecionadas via iPCA), consideramos a possibilidade de haver adulteração de algumas dessas amostras com solventes aromáticos. Este tipo de adulterante tem sido estudado na gasolina e diesel.^{52,54,56–60} Kulathunga e Mahanama⁵⁷ usaram as características de fluorescência dos PAHs para estudar a adulteração em misturas sintéticas em vinte amostras comerciais de diesel e gasolina. Eles usaram o método de espectroscopia de fluorescência síncrona, uma vez que os PAHs têm alta sensibilidade à fluorescência com suas propriedades de fluorescência molecular pelo sistema de elétrons π nas estruturas de anéis fundidos. Eles encontraram duas amostras com maiores níveis de adulteração (25 e 38 %v/v) e duas com 12 %v/v, de um total de dez amostras comerciais de diesel. Os compostos de PAHs presentes no problema da adulteração do diesel também foram avaliados com diferentes substâncias, como querosene, terebentina e óleos residuais.^{54,58} Além disso, Babu et al.⁵⁹ relacionaram o potencial de seis solventes orgânicos comerciais como adulterante para a base de diesel em corporações petrolíferas indianas. Um desses adulterantes é um solvente composto de hidrocarbonetos aromáticos. Os outros cinco são misturas de compostos alifáticos e aromáticos.

Da mesma forma, outros estudos relataram que amostras adulteradas podem não ser detectadas por meio de parâmetros físico-químicos.^{4,60-62} Vempatapu et al.⁶¹ avaliaram os parâmetros físico-químicos do diesel misturado com querosene e concluíram que alguns parâmetros como densidade, viscosidade cinemática e índice de cetano apenas revelam níveis elevados de adulteração. A partir de dezesseis amostras de diesel, Tosato et al.⁶⁰ encontraram amostras com maior teor aromático pela análise de ressonância magnética nuclear de hidrogênio (RMN de ^1H), embora não tenham encontrado não conformidades de acordo com a legislação brasileira.

Duas das amostras sugeridas pela URF também foram sugeridas como não conformes por meio de modelos SIMCA distintos, conduzidos a partir das mesmas regiões espectrais escolhidas pela iPCA, compreendendo carbonila ($1816\text{-}1650\text{ cm}^{-1}$) e aromáticos ($1650\text{-}1500\text{ cm}^{-1}$; $1300\text{-}1150\text{ cm}^{-1}$; e $950\text{-}650\text{ cm}^{-1}$). SIMCA apresentou desempenho diferente de acordo com as regiões espectroscópicas (Tabela S1 a Tabela S4). Como esperado, a região carbonila fornece o modelo com maior desempenho (96% e 86% de exatidão, para treinamento e teste, respectivamente) enquanto a região da impressão digital teve um desempenho inferior (63% e 67% de exatidão, para treinamento e teste, respectivamente).

Da mesma forma, Ranzan et al.¹⁵ encontraram uma diferença qualitativa significativa entre o diesel S10 e S100 a partir das medições de espectroscopia de fluorescência 2D pelos modelos PCA, RF e PLS-DA, apresentando desempenho semelhante. Neste trabalho, propusemos uma triagem a partir de um método simples e não supervisionado para identificar informações que possam sugerir adulteração das amostras comerciais de diesel recebidas. A partir desse modelo único não supervisionado pelo método URF, as amostras duvidosas puderam ser identificadas. Enquanto pelos modelos PCA e SIMCA foi preciso estratificar os espectros para chegar à mesma informação. O método URF é uma excelente alternativa para testar fraudes por espectroscopia a partir de um conjunto de amostras pequeno e não rotulado. A amostragem empregada no algoritmo de RF com subconjuntos de variáveis preditoras por meio do parâmetro '*mtry*', reduz o espaço avaliado em cada nó. Este parâmetro amostrou aleatoriamente variáveis que são avaliadas como possíveis preditoras. Além disso, apesar de um menor número de árvores reduzir o custo do processamento de máquina, neste estudo, um teste prévio com diferentes números de árvores mostrou que uma quantidade suficientemente grande (1000)

garantiu a exatidão necessária e a relação entre as variáveis colineares adequadamente capturadas pelo procedimento *In-Bag*.^{18-22,43}

3.4. Conclusões

O uso de floresta aleatória não supervisionada possibilitou discriminar amostras de diesel que também apresentaram não conformidade quanto à acidez e ponto de fulgor por meio de parâmetros físico-químicos. Além disso, algumas amostras apresentaram alterações nos espectros MIR nas regiões de 1600 cm^{-1} e $900\text{-}650\text{ cm}^{-1}$, foram destacadas no modelo URF e nos modelos PCA e SIMCA dessas regiões destacadas. Essas alterações podem sugerir adulteração com solvente aromático, o que não foi evidenciado pelos parâmetros físico-químicos.

Dados espectroscópicos associados à análise multivariada não supervisionada e supervisionada são métodos alternativos importantes para o rastreamento rápido e preciso neste problema. As respostas do iPCA e do SIMCA são coerentes com os testes físico-químicos na região específica dos espectros MIR ($1816\text{ a }1650\text{ cm}^{-1}$). No entanto, o *ensemble* URF pode sugerir amostras com similaridade a adulterações forjadas, mesmo que essas amostras não sejam identificadas por meio de parâmetros físico-químicos ou modelos PCA. O método URF apresenta algumas vantagens, como as relações lineares e não lineares sendo modeladas, com um consequente alcance de informações que talvez na PCA não seja possível. Além disso, ele fornece a matriz de proximidade e uma visualização no espaço de baixa dimensão por meio do gráfico PcoA, que pode ser interpretado conforme é realizado pela PCA. Assim, o modelo de URF pode ser um método de triagem de adulteração.

Agradecimentos

Os autores agradecem ao NCQP/LABPETRO-UFES, FAPES, CAPES e CNPq.

Financiamento: Este trabalho foi apoiado pela CAPES (Código Financeiro 001); FAPES (processo número 33530.503.20537.12092017; nº 151/2020, nº 165/2021 e nº 691/2022) e CNPq (processo nº 422515/2016-7; #401870/2020-0 e #310057/2020-5).

3.5. Referências

- (1) Yadav, Sh. R.; Murthy, V. K.; Mishra, D.; Baral, B. Estimation of Petrol and Diesel Adulteration with Kerosene and Assessment of Usefulness of Selected Automobile Fuel Quality Test Parameters. *International Journal of Environmental Science & Technology* **2005**, 1 (4), 253–255. <https://doi.org/10.1007/bf03325839>.
- (2) Gedik, K.; Uzun, Y. Characterization of the Properties of Diesel-Base Oil-Solvent-Waste Oil Blends Used as Generic Fuel in Diesel Engines. *Fuel Processing Technology* **2015**, 139 (x), 135–141. <https://doi.org/10.1016/j.fuproc.2015.07.032>.
- (3) Gontijo, L. C.; Guimarães, E.; Mitsutake, H.; Santana, F. B. de; Santos, D. Q.; Borges Neto, W. Quantification of Soybean Biodiesels in Diesel Blends According to ASTM E1655 Using Mid-Infrared Spectroscopy and Multivariate Calibration. *Fuel* **2014**, 117 (PARTB), 1111–1114. <https://doi.org/10.1016/j.fuel.2013.10.043>.
- (4) Moura, H. O. M. A.; Câmara, A. B. F.; Santos, M. C. D.; Morais, C. L. M.; de Lima, L. A. S.; Lima, K. M. G.; de Carvalho, L. S. Advances in Chemometric Control of Commercial Diesel Adulteration by Kerosene Using IR Spectroscopy. *Anal Bioanal Chem* **2019**, 411 (11), 2301–2315. <https://doi.org/10.1007/s00216-019-01671-y>.
- (5) Parsons, B. A.; Pinkerton, D. K.; Wright, B. W.; Synovec, R. E. Chemical Characterization of the Acid Alteration of Diesel Fuel: Non-Targeted Analysis by Two-Dimensional Gas Chromatography Coupled with Time-of-Flight Mass Spectrometry with Tile-Based Fisher Ratio and Combinatorial Threshold Determination. *J Chromatogr A* **2016**, 1440, 179–190. <https://doi.org/10.1016/j.chroma.2016.02.067>.
- (6) Palou, A.; Miró, A.; Blanco, M.; Larraz, R.; Gómez, J. F.; Martínez, T.; González, J. M.; Alcalà, M. Calibration Sets Selection Strategy for the Construction of Robust PLS Models for Prediction of Biodiesel/Diesel Blends Physico-Chemical Properties Using NIR Spectroscopy. *Spectrochim Acta A Mol Biomol Spectrosc* **2017**, 180, 119–126. <https://doi.org/10.1016/j.saa.2017.03.008>.
- (7) Aleme, H. G.; Corgozinho, C. N. C.; Barbeira, P. J. S. Diesel Oil Discrimination by Origin and Type Using Physicochemical Properties and Multivariate Analysis. *Fuel* **2010**, 89 (11), 3151–3156. <https://doi.org/10.1016/j.fuel.2010.05.010>.
- (8) Bassbasi, M.; Hafid, A.; Platikanov, S.; Tauler, R.; Oussama, A. Study of Motor Oil Adulteration by Infrared Spectroscopy and Chemometrics Methods. *Fuel* **2013**, 104, 798–804. <https://doi.org/10.1016/j.fuel.2012.05.058>.
- (9) Ferrão, M. F.; Viera, M. D. S.; Pazos, R. E. P.; Fachini, D.; Gerbase, A. E.; Marder, L. Simultaneous Determination of Quality Parameters of Biodiesel/Diesel Blends Using HATR-FTIR Spectra and PLS, IPLS or SiPLS Regressions. *Fuel* **2011**, 90 (2), 701–706. <https://doi.org/10.1016/j.fuel.2010.09.016>.
- (10) Mazivila, S. J.; Olivieri, A. C. Chemometrics Coupled to Vibrational Spectroscopy and Spectroscopic Imaging for the Analysis of Solid-Phase Pharmaceutical Products: A Brief Review on Non-Destructive Analytical Methods. *TrAC - Trends in Analytical Chemistry* **2018**, 108, 74–87. <https://doi.org/10.1016/j.trac.2018.08.013>.
- (11) Obeidat, S. M.; Al-Ktash, M. M.; Al-Momani, I. F. Study of Fuel Assessment and Adulteration Using EEMF and Multiway PCA. *Energy and Fuels* **2014**, 28 (8), 4889–4894. <https://doi.org/10.1021/ef500718e>.

- (12) Câmara, A. B. F.; de Carvalho, L. S.; de Moraes, C. L. M.; de Lima, L. A. S.; de Araújo, H. O. M.; de Oliveira, F. M.; de Lima, K. M. G. MCR-ALS and PLS Coupled to NIR/MIR Spectroscopies for Quantification and Identification of Adulterant in Biodiesel-Diesel Blends. *Fuel* **2017**, *210* (July), 497–506. <https://doi.org/10.1016/j.fuel.2017.08.072>.
- (13) de Vasconcelos, F. V. C.; de Souza, P. F. B.; Pimentel, M. F.; Pontes, M. J. C.; Pereira, C. F. Using Near-Infrared Overtone Regions to Determine Biodiesel Content and Adulteration of Diesel/Biodiesel Blends with Vegetable Oils. *Anal Chim Acta* **2012**, *716*, 101–107. <https://doi.org/10.1016/j.aca.2011.12.027>.
- (14) Aleme, H. G.; Corgozinho, C. N. C.; Barbeira, P. J. S. Diesel Oil Discrimination by Origin and Type Using Physicochemical Properties and Multivariate Analysis. *Fuel* **2010**, *89* (11), 3151–3156. <https://doi.org/10.1016/j.fuel.2010.05.010>.
- (15) Ranzan, L.; Ranzan, C.; Trierweiler, L. F.; Trierweiler, J. O. Classification of Diesel Fuel Using Two-Dimensional Fluorescence Spectroscopy. *Energy and Fuels* **2017**, *31* (9), 8942–8950. <https://doi.org/10.1021/acs.energyfuels.7b00954>.
- (16) Miraboutalebi, S. M.; Kazemi, P.; Bahrami, P. Fatty Acid Methyl Ester (FAME) Composition Used for Estimation of Biodiesel Cetane Number Employing Random Forest and Artificial Neural Networks: A New Approach. *Fuel* **2016**, *166*, 143–151. <https://doi.org/10.1016/j.fuel.2015.10.118>.
- (17) Le, B. T.; Ha, T. T. L. Total Aromatics of Diesel Fuels Analysis by Deep Learning and Near-Infrared Spectroscopy. *Spectroscopy Letters* **2019**, *52* (10), 671–676. <https://doi.org/10.1080/00387010.2019.1682014>.
- (18) Lovatti, B. P. O.; Nascimento, M. H. C.; Neto, Á. C.; Castro, E. V. R.; Filgueiras, P. R. Use of Random Forest in the Identification of Important Variables. *Microchem. J.* **2019**, *145* (December 2018), 1129–1134. <https://doi.org/10.1016/j.microc.2018.12.028>.
- (19) Oliveira, B. P. L.; Nascimento, M. H. C.; Rainha, K. P.; Alvaro, E. C. S. O.; Filgueiras, P. R. Different Strategies for the Use of Random Forest in NMR Spectra. *J. Chemometrics* **2020**, No. Special issue, 1–10. <https://doi.org/10.1002/cem.3231>.
- (20) Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45* (1), 5–32. <https://doi.org/10.1023/A:101093340>.
- (21) Breiman, L. Bagging Predictors. *Mach. Learn.* **1996**, *24* (2), 123–140. <https://doi.org/10.1007/bf00058655>.
- (22) Afanador, N. L.; Smolinska, A.; Tran, T. N.; Blanchet, L. Unsupervised Random Forest: A Tutorial with Case Studies. *J. Chemometrics* **2016**, *30* (5), 232–241. <https://doi.org/10.1002/cem.2790>.
- (23) Chen, S.; Plouffe, A.; Hattori, K. A Multivariate Statistical Approach Identifying the Areas Underlain by Potential Porphyry-Style Cu Mineralization, South-Central British Columbia, Canada. *J Geochem Explor* **2019**, *202*, 13–26. <https://doi.org/10.1016/j.gexplo.2019.03.016>.
- (24) Poulakis, K.; Pereira, J. B.; Mecocci, P.; Vellas, B.; Tsolaki, M.; Kłoszewska, I.; Soininen, H.; Lovestone, S.; Simmons, A.; Wahlund, L. O.; Westman, E.

- Heterogeneous Patterns of Brain Atrophy in Alzheimer's Disease. *Neurobiol Aging* **2018**, 65, 98–108. <https://doi.org/10.1016/j.neurobiolaging.2018.01.009>.
- (25) Baron, D.; Poznanski, D. The Weirdest SDSS Galaxies: Results from an Outlier Detection Algorithm. *Mon Not R Astron Soc* **2017**, 465 (4), 4530–4555. <https://doi.org/10.1093/mnras/stw3021>.
 - (26) Reis, I.; Poznanski, D.; Baron, D.; Zasowski, G.; Shahaf, S. Detecting Outliers and Learning Complex Structures with Large Spectroscopic Surveys - A Case Study with APOGEE Stars. *Mon Not R Astron Soc* **2018**, 476 (2), 2117–2136. <https://doi.org/10.1093/mnras/sty348>.
 - (27) Peerbhay, K.; Mutanga, O.; Lottering, R.; Ismail, R. Mapping Solanum Mauritianum Plant Invasions Using WorldView-2 Imagery and Unsupervised Random Forests. *Remote Sens Environ* **2016**, 182, 39–48. <https://doi.org/10.1016/j.rse.2016.04.025>.
 - (28) de Maat, M. F. G.; van de Velde, C. J. H.; Benard, A.; Putter, H.; Morreau, H.; van Krieken, J. H. J. M.; Klein-Kranenburg, E. M.; de Graaf, E. J.; Tollenaar, R. A. E. M.; Hoon, D. S. B. Identification of a Quantitative MINT Locus Methylation Profile Predicting Local Regional Recurrence of Rectal Cancer. *Clinical Cancer Research* **2010**, 16 (10), 2811–2818. <https://doi.org/10.1158/1078-0432.CCR-09-2717>.
 - (29) Brasil. Agência Nacional de Petróleo Gás Natural e Biocombustíveis. *Resolução ANP Nº 52 (RANP 52-2010)*; 2010.
 - (30) Brasil. Agência Nacional de Petróleo Gás Natural e Biocombustíveis. *Resolução ANP Nº 50, DE 23.12.2013. Agência Nacional do Petróleo, Gás Natural e Biocombustíveis* **2013**, 11–13.
 - (31) ASTM International: West Conshohocken. ASTM D4052-18a - Standard Test Method for Density , Relative Density , and API Gravity of Liquids by Digital Density Meter. *Annual Book of ASTM Standards* **2013**, 1–8. <https://doi.org/10.1520/D4052-18A.2>.
 - (32) ASTM International: West Conshohocken. ASTM D93-18 - Standard Test Methods for Flash Point by Pensky-Martens Closed Cup Tester. *Annual Book of ASTM Standards* **2007**, i (Reapproved 2006), 1–5. <https://doi.org/10.1520/D0093-18>.
 - (33) ASTM International: West Conshohocken. ASTM D664-18 - Standard Test Method for Acid Number of Petroleum Products by Potentiometric Titration,. *Annual Book of ASTM Standards* **2018**, 1–16. <https://doi.org/10.1520/D0664-18E02.Este>.
 - (34) ASTM International: West Conshohocken. ASTM D4294-16 - Standard Test Method for Sulfur in Petroleum and Petroleum Products by Energy Dispersive X-Ray Fluorescence Spectrometry. *Annual Book of ASTM Standards* **2016**, 552 (c), 1–9. <https://doi.org/10.1520/D4294-16E01.2>.
 - (35) ASTM. Standard D7042: Test Method for Dynamic Viscosity and Density of Liquids by Stabinger Viscometer (and the Calculation of Kinematic Viscosity). *American National Standard Institute* **2013**, 12a, 1–11. <https://doi.org/10.1520/D7042-16E03.priate>.
 - (36) ASTM International: West Conshohocken. ASTM D97-17b - Standard Test Method for Pour Point of Petroleum Products. *Annual Book of ASTM Standards* **2017**, 1–7. <https://doi.org/10.1520/D0097-17B.In>.

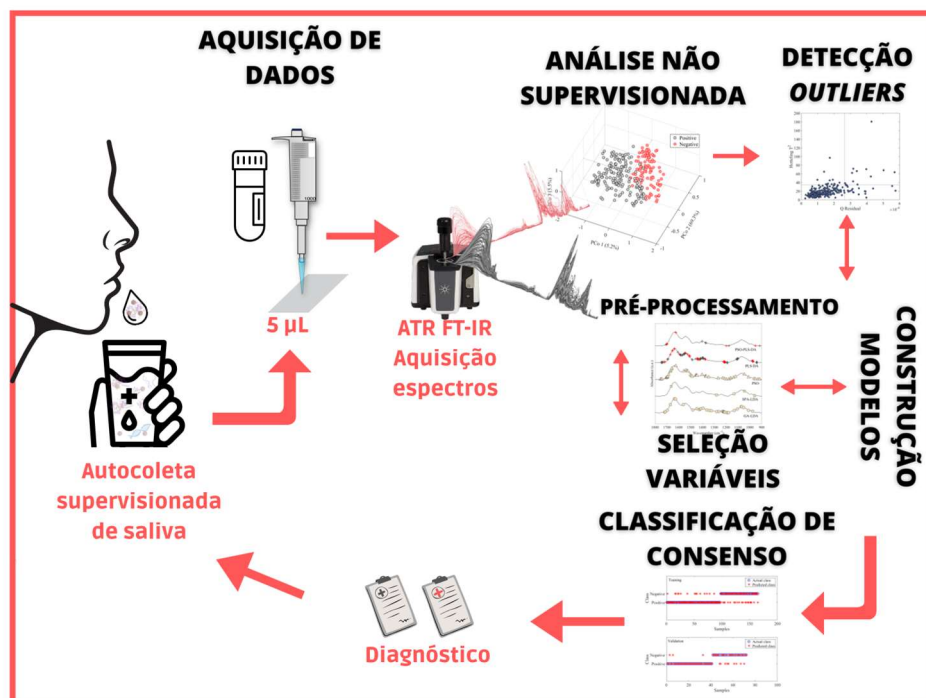
- (37) Savitzky, A.; Golay, M. J. E. Smoothing and Differentiation of Data by Simplified Least Squares Procedures. *Anal. Chem.* **1964**, *36* (8), 1627–1639. <https://doi.org/10.1021/ac60214a047>.
- (38) Barnes, R. J.; Dhanoa, M. S.; Lister, S. J. Standard Normal Variate Transformation and De-Trending of near-Infrared Diffuse Reflectance Spectra. *Appl. Spectroscopy* **1989**, *43* (5), 772–777. <https://doi.org/10.1366/0003702894202201>.
- (39) Isaksson, T.; Naes, T. Effect of Multiplicative Scatter Correction (MSC) and Linearity Improvement in NIR Spectroscopy. *Appl. Spectroscopy* **1988**, *42* (7), 1273–1284. <https://doi.org/10.1366/0003702884429869>.
- (40) Nørgaard, L.; Saudland, A.; Wagner, J.; Nielsen, J. P.; Munck, L.; Engelsen, S. B. Interval Partial Least-Squares Regression (IPLS): A Comparative Chemometric Study with an Example from near-Infrared Spectroscopy. *Appl Spectrosc* **2000**, *54* (3), 413–419. <https://doi.org/10.1366/0003702001949500>.
- (41) Wold, S.; Esbensen, K.; Geladi, P. Principal Component Analysis. *Chemometrics and Intelligent Laboratory Systems* **1987**, *2* (1–3), 37–52. [https://doi.org/10.1016/0169-7439\(87\)80084-9](https://doi.org/10.1016/0169-7439(87)80084-9).
- (42) Bro, R.; Smilde, A. K. Principal Component Analysis. *Anal. Methods* **2014**, *6*, 2812–2831. <https://doi.org/10.1039/c3ay41907j>.
- (43) Lee, S.; Choi, H.; Cha, K.; Chung, H. Random Forest as a Potential Multivariate Method for Near-Infrared (NIR) Spectroscopic Analysis of Complex Mixture Samples: Gasoline and Naphtha. *Microchemical Journal* **2013**, *110*, 739–748. <https://doi.org/10.1016/j.microc.2013.08.007>.
- (44) Shi T; S., H. Unsupervised Learning with Random Forest Predictors. *Journal of Computational and Graphical Statistics. Graphical Stat.* **2006**, *15* (1), 118–138.
- (45) Wold, S. Pattern Recognition by Means of Disjoint Principal Components Models. *Pattern Recognit* **1976**, *8* (3), 127–139. [https://doi.org/10.1016/0031-3203\(76\)90014-5](https://doi.org/10.1016/0031-3203(76)90014-5).
- (46) Vitale, R.; Marini, F.; Ruckebusch, C. SIMCA Modeling for Overlapping Classes: Fixed or Optimized Decision Threshold? *Anal Chem* **2018**, *90* (18), 10738–10747. <https://doi.org/10.1021/acs.analchem.8b01270>.
- (47) Ballabio, D.; Consonni, V. Classification Tools in Chemistry. Part 1: Linear Models. PLS-DA. *Anal. Methods* **2013**, *5* (16), 3790–3798. <https://doi.org/10.1039/c3ay40582f>.
- (48) Brereton, R. G. Contingency Tables, Confusion Matrices, Classifiers and Quality of Prediction. *J. Chemometrics* **2021**, e3331, 1–5. <https://doi.org/10.1002/cem.3331>.
- (49) Ballabio, D.; Grisoni, F.; Todeschini, R. Multivariate Comparison of Classification Performance Measures. *Chemometrics and Intelligent Laboratory Systems* **2018**, *174* (April 2017), 33–44. <https://doi.org/10.1016/j.chemolab.2017.12.004>.
- (50) Nespeca, M. G.; Hatanaka, R. R.; Flumignan, D. L.; de Oliveira, J. E. Rapid and Simultaneous Prediction of Eight Diesel Quality Parameters through ATR-FTIR Analysis. *J Anal Methods Chem* **2018**, *2018*. <https://doi.org/10.1155/2018/1795624>.

- (51) Silverstein, R. M.; Webster, F. X.; Kiemle, D. J.; Bryce, D. L. *Spectrometric Identification of Organic Compounds*, 8th ed.; John Wiley & Sons: New York, 2014.
- (52) Vempatapu, B. P.; Kanaujia, P. K. Monitoring Petroleum Fuel Adulteration: A Review of Analytical Methods. *TrAC - Trends in Analytical Chemistry* **2017**, 92, 1–11. <https://doi.org/10.1016/j.trac.2017.04.011>.
- (53) Mazivila, S. J.; Gontijo, L. C.; Santana, F. B. de; Mitsutake, H.; Santos, D. Q.; Borges Neto, W. Fast Detection of Adulterants/Contaminants in Biodiesel/Diesel Blend (B5) Employing Mid-Infrared Spectroscopy and PLS-DA. *Energy and Fuels* **2015**, 29 (1), 227–232. <https://doi.org/10.1021/ef502122w>.
- (54) Corgozinho, C. N. C.; Pasa, V. M. D.; Barbeira, P. J. S. Determination of Residual Oil in Diesel Oil by Spectrofluorimetric and Chemometric Analysis. *Talanta* **2008**, 76 (2), 479–484. <https://doi.org/10.1016/j.talanta.2008.03.003>.
- (55) Fernanda Pimentel, M.; Ribeiro, G. M. G. S.; da Cruz, R. S.; Stragevitch, L.; Pacheco Filho, J. G. A.; Teixeira, L. S. G. Determination of Biodiesel Content When Blended with Mineral Diesel Fuel Using Infrared Spectroscopy and Multivariate Calibration. *Microchemical Journal* **2006**, 82 (2), 201–206. <https://doi.org/10.1016/j.microc.2006.01.019>.
- (56) Bezerra, A. C. de M.; Silva, D. de O.; Matos, G. H. M. de; dos Santos, J. P.; Borges, C. N.; Silveira, L.; Pacheco, M. T. T. Quantification of Anhydrous Ethanol and Detection of Adulterants in Commercial Brazilian Gasoline by Raman Spectroscopy. *Instrum Sci Technol* **2019**, 47 (1), 90–106. <https://doi.org/10.1080/10739149.2018.1470535>.
- (57) Kulathunga, D. R.; Mahanama, K. R. R. Fingerprinting Diesel and Petrol Fuels for Adulteration in Sri Lanka. *J Natl Sci Found* **2013**, 41 (4), 287–292. <https://doi.org/10.4038/jnsfsr.v41i4.6247>.
- (58) Meira, M.; Quintella, C. M.; Ribeiro, E. M. O.; Silva, H. R. G.; Guimarães, A. K.; Luna, S.; Silva, W. L.; de Brito, I. J. Determination of Adulterants in Diesel by Integration of LED Fluorescence Spectra. *J Braz Chem Soc* **2015**, 26 (7), 1351–1356. <https://doi.org/10.5935/0103-5053.20150103>.
- (59) Babu, V.; Krishna, R.; Mani, N. Review on the Detection of Adulteration in Fuels through Computational Techniques. *Mater Today Proc* **2017**, 4 (2), 1723–1729. <https://doi.org/10.1016/j.matpr.2017.02.013>.
- (60) Tosato, F.; Barros, E.; Cunha, D.; Santos, F.; Corrêa, T.; Nunes, A.; Jastrow, I.; Silverol, M.; Pinheiro, L.; Seabra, A.; Filgueiras, P.; Neto, Á.; Wang, L.; Ferreira, E.; Romão, W. Análise De Amostras De Combustíveis Por Fotometria, Nir Portátil E Rmn De 1 H – Uma Comparação Com Os Resultados Encontrados Por Técnicas Normatizadas. *Quim Nova* **2020**, X (00), 1–13. <https://doi.org/10.21577/0100-4042.20170473>.
- (61) Vempatapu, B. P.; Tripathi, D.; Kumar, J.; Kanaujia, P. K. Determination of Kerosene as an Adulterant in Diesel through Chromatography and High-Resolution Mass Spectrometry. *SN Appl Sci* **2019**, 1 (6), 1–12. <https://doi.org/10.1007/s42452-019-0637-7>.

- (62) Kalligeros, S.; Zannikos, F.; Stournas, S.; Lois, E. Fuel Adulteration Issues in Greece. *Energy* **2003**, 28 (1), 15–26. [https://doi.org/10.1016/S0360-5442\(02\)00091-9](https://doi.org/10.1016/S0360-5442(02)00091-9).

CAPÍTULO 4

DIAGNÓSTICO NÃO INVASIVO PARA COVID-19 VIA ESPECTROSCOPIA FTIR E ANÁLISE MULTIVARIADA A PARTIR DE SALIVA



- Neste capítulo, o método de floresta aleatória não supervisionada teve uma participação importante na identificação de estrutura nos dados de espectroscopia na região do infravermelho médio de amostras de saliva de pacientes acometidos pela infecção COVID-19, em comparação com pacientes com sintomas semelhantes, mas de diagnóstico negativo para COVID-19.
- Uma classificação por consenso utilizando análise discriminante com diferentes seleções de variáveis, distinguiu amostras de pacientes positivos por meio da espectroscopia na região do infravermelho médio com sensibilidade de 93% e especificidade de 83%.

4. DIAGNÓSTICO NÃO INVASIVO PARA COVID-19 VIA ESPECTROSCOPIA FTIR E ANÁLISE MULTIVARIADA A PARTIR DE SALIVA

Resumo

A síndrome respiratória aguda grave coronavírus 2 (SARS-CoV-2) tem ocasionado a pior crise de saúde global do século. A reação em cadeia de polimerase de transcrição reversa (RT-qPCR) é o método de diagnóstico padrão-ouro, mas apresenta limitações frente à grande demanda. Avaliamos um conjunto de dados de infravermelho médio (MIR) de 237 amostras de saliva de pacientes sintomáticos (138 infecções diagnosticadas como covid-19 pelo RT-qPCR). Os espectros MIR foram avaliados por meio de floresta aleatória não supervisionada (URF) e modelos de classificação. A análise linear discriminante (LDA) foi aplicada por meio de algoritmo genético (GA-LDA), algoritmo de projeção sucessiva (SPA-LDA), mínimos quadrados parciais (PLS-DA) e uma combinação de métodos de redução de dimensões e seleção variável por otimização de enxame de partículas (PSO-PLSDA). Além disso, foi aplicada classificação de consenso. O modelo URF pode identificar uma estrutura em dados mesmo sob alta complexidade. Os modelos individuais apresentaram bom desempenho, mas a classificação de consenso melhorou o desempenho de validação com exatidão de 85%, sensibilidade de 93%, especificidade de 83% e coeficiente de correlação de Matthews de 0,69, com informações de diferentes regiões espectrais. Portanto, por meio destes métodos propostos não supervisionados e supervisionados é possível destacar regiões espectrais de maior relação com amostras positivas, como lipídios ($\sim 1700\text{ cm}^{-1}$), proteínas ($\sim 1400\text{ cm}^{-1}$) e regiões de ácido nucleico ($\sim 1200 - 950\text{ cm}^{-1}$). Esta metodologia apresenta uma importante ferramenta para uma técnica de diagnóstico, não invasiva, em pouco tempo, reduzindo custos e riscos.

Palavras-chave: COVID-19; MIR; Saliva; quimiometria; floresta aleatória não supervisionada; otimização por enxame de partículas.

4.1. Introdução

A pandemia causada por síndrome respiratória aguda grave - coronavírus 2 (SARS-CoV-2) tem afetado países ao redor do mundo desde sua incidência na cidade de Wuhan, China, em dezembro de 2019. Globalmente, segundo a Organização Mundial da Saúde (OMS), até 1º de julho de 2021, houve 181.930,736 casos confirmados de doença coronavírus 2019 (COVID-19), incluindo 3.945,832 óbitos.¹ Uma das ações críticas para controlar a propagação do vírus é isolar rapidamente as pessoas infectadas. Para isso, há necessidade de detecção de vírus em larga escala com precisão, velocidade e confiabilidade.¹⁻⁵ Os ensaios de detecção de vírus mais conhecidos são o ensaio imunoenzimático (ELISA, do inglês *Enzyme-Linked Immunosorbent Assay*) e reação em cadeia de polimerase quantitativa de transcrição reversa (RT-qPCR). Este último tem sido usado como padrão-ouro para o diagnóstico de infecção SARS-CoV-2. Esses métodos apresentam alta repetibilidade, reprodutibilidade e robustez.⁶ No entanto, eles requerem recursos laboratoriais e reagentes químicos. Além disso, o tempo de resposta, a logística das amostras e outros fatores devem ser avaliados devido à grande demanda nas condições de pandemia. Assim, é urgente desenvolver métodos confiáveis e rápidos para atender à demanda em diagnósticos em larga escala para enfrentar a pandemia atual.^{1-4,7}

Neste sentido, técnicas de espectroscopia vibracional como o infravermelho (IR) têm sido discutidas como de grande potencial, uma vez que são reprodutíveis, não invasivas, e precisam de pouco ou nenhum preparo amostral e outros reagentes. Além disso, as informações em níveis moleculares em que inferem sobre grupos funcionais, tipos de vínculos e conformação molecular, podem identificar importantes alterações bioquímicas em amostras biológicas na presença de vírus.^{6,8} A espectroscopia de IR é uma das principais técnicas de avaliação de modos vibracionais moleculares, baseada em mudanças no momento do dipolo causadas por vibrações das ligações químicas. Esses movimentos vibracionais permitem que as moléculas absorvam energia de radiação relacionada ao seu nível de energia vibracional. Entre as regiões de IR, o IR na região do médio (MIR), com faixa de número de ondas de 4.000 a 200 cm^{-1} , tem sido o principal interesse na análise biológica, pois incluem regiões de absorção de importantes biomoléculas, principalmente na região de 1.800 a 900 cm^{-1} que é conhecida como *biofingerprint*,

associada à estrutura e função dos biofluidos analisados.⁹ Nesta região, existem absorções de diferentes constituintes, biomoléculas que podem ser biomarcadores como lipídios (alongamento simétrico C=O ($\sim 1.750\text{ cm}^{-1}$), carboidratos (alongamento simétrico CO-O-C a $\sim 1.155\text{ cm}^{-1}$), ácido nucleico (estiramento assimétrico de grupo fosfato, $\sim 1.225\text{ cm}^{-1}$, e simétrico $\sim 1.080\text{ cm}^{-1}$), além de glicogênio e fosforilação de proteína (entre ~ 1.030 e 900 cm^{-1}), e outros. Proteínas têm maiores contribuições de sinal de amida I em $\sim 1.650\text{ cm}^{-1}$ (estiramento de C=O, deformação angular C-N e estiramento C-N) e amida II a $\sim 1.550\text{ cm}^{-1}$ (deformação angular de N-H e estiramento C-N), e menor contribuição de amida III em $\sim 1.260\text{ cm}^{-1}$ (estiramento C-N).^{6,8,10–12}

Estudos têm sido relatados com uso de MIR para detectar o vírus da dengue em amostras de sangue,¹³ identificação de bactérias *staphylococcus aureus* em amostras de sangue,¹⁴ estudo de estabilidade da composição sanguínea de pessoas saudáveis¹⁵ e outros. No entanto, uma das dificuldades no uso da espectroscopia MIR em estudos de virologia está relacionada à variabilidade dos vírus existentes em organismos humanos. Uma maior variabilidade do vírus implica sobreposição de informações espectrais, além de toda a especificidade do organismo humano que geram amostras complexas. Para resolver esses desafios, estudos de biospectroscopia têm sido associados a métodos estatísticos de aprendizagem. Os principais métodos aplicados são a análise de componentes principais (PCA), análise hierárquica de cluster (HCA) e análise discriminante linear (LDA). Este último requer métodos de redução de dimensão ou seleção de variáveis, devido aos dados de espectroscopia que geralmente apresentam variáveis com colinearidade. Assim, ao lado do LDA estão aplicados algoritmo genético (GA-LDA), algoritmo de projeção sucessiva (SPA-LDA), PCA-DA e mínimos quadrados parciais com análise discriminante (PLS-DA).^{6,8,10–14,16–18}

A partir de 2020, alguns grupos de pesquisa têm voltado sua atenção para o uso de métodos de espectroscopia para detectar o vírus SARS-CoV-2, como desenvolvimento de biosensores plasmônicos,¹⁹ detecção de vírus em amostras de sangue,²⁰ em esfregaços de células orais ou faríngeas.²¹ Baraúna *et al.*²¹ analisaram esfregaços de células orais e faríngeas por meio de coleta de *swab* de pacientes com sintomas semelhantes a infecções COVID-19. As amostras foram separadas em treinamento (50 positivos, 50 negativos) e validação (20 positivos, 61 negativos) e classificadas pelo GA-LDA com sensibilidade de 95% e especificidade de 89%. No

entanto, associaram apenas uma das cinco variáveis selecionadas com o RNA do vírus e outras variáveis selecionadas com resposta inflamatória do organismo. Wood *et al.*⁴ estudaram sinais espectroscópicos característicos de biomarcadores SARS-CoV-2 com espectros sincrotron-FTIR e Raman do vírus purificado. Para o diagnóstico de infecção COVID-19, modelaram 57 espectros médios, dos quais 29 são positivos para infecção por SARS-CoV-2 por RT-qPCR, e 28 negativos. Com espectros truncados de 1.300-900 cm^{-1} e PLSDA com um limiar otimizado de 0,6, obtiveram sensibilidade de 93% e especificidade de 82%. No entanto, concluíram que precisam de um maior conjunto de pacientes para melhorar a sensibilidade e a especificidade e avaliar se esta técnica distingue outros vírus respiratórios.

Até onde é de nosso conhecimento, não há estudos com diagnóstico por meio de espectroscopia de saliva de um coorte grande de pacientes. Nesse sentido, avaliamos o uso de espectroscopia MIR associada a métodos de reconhecimento de padrões para diagnosticar um maior número de pacientes a partir de saliva, classificando entre amostras positivas ou negativas a infecções por SARS-CoV-2. Este trabalho visa contribuir na busca por métodos de diagnósticos rápidos e menos invasivos como alternativa ao rastreamento de pacientes com sintomas suspeitos de infecção pela COVID-19.

4.2. Metodologia

4.2.1 Pacientes Participantes

Neste estudo, foram avaliados espectros MIR de um total de 265 pacientes assistidos pelos serviços de saúde do Espírito Santo. Esses pacientes foram atendidos de acordo com as diretrizes das Secretarias Estaduais de Saúde e, segundo a OMS. Este estudo foi realizado de acordo com a declaração de Helsinque e autorizado pela Diretiva hospitalar devido à situação de emergência. A aprovação ética foi concedida pelo Comitê de Ética da Universidade Federal do Espírito Santo (nº 0993920.1.0000.5071 e nº 31411420.9.0000.8207). Todos os pacientes concordaram em participar voluntariamente e foram esclarecidos por meio de termo de esclarecimento e consentimento. Em seguida, o *swab* nasofaríngeo foi coletado por meio de um profissional de saúde para análise RT-qPCR. Em seguida, o paciente recebeu um tubo estéril para autocoleta orientada de saliva. Todas as etapas desde

a admissão dos participantes até os modelos de classificação estão descritas na Figura 4.1.



Figura 4.1 Diagrama descritivo da metodologia deste estudo, contendo as etapas de admissão dos pacientes e coleta do biofluidos, aquisição dos dados e análise multivariada dos dados.

As análises RT-qPCR, foram realizadas no laboratório Central da Secretaria de Estado da Saúde do Espírito Santo (LACEN-SESA). Estes resultados de RT-qPCR foram utilizados para uma atribuição de classe para amostras, gerando um

vetor de resposta de classe para treinamento do modelo de aprendizagem de máquina. O Laboratório utilizou um tipo de kit de diagnóstico molecular com limite de detecção $\leq 6,363$ cópias/reação para gene alvo N e 18,705 cópias/reação para gene alvo ORF1ab. Além disso, o limite de detecção foi apresentado para cópias do vírus/ μL de amostra (μL para cada reação): 1,27 cópias/ μL para gene N e 3,74 μL para gene ORF1ab. As características de desempenho, sensibilidade e especificidade clínica, foram avaliadas com um teste contendo 240 amostras (163 negativas e 77 positivas), com valores de: 92% de concordância; 100% de sensibilidade e 88% de especificidade.

4.2.2 Espectroscopia MIR

Para análise espectral, 5 μL de saliva foram transferidos para um recorte de papel alumínio para procedimento de secagem. Este procedimento foi adotado para minimizar a interferência da água (componente majoritário na composição química da saliva) na aquisição dos espectros. As amostras permaneceram em sala com controle de umidade ($\sim 40\%$) e temperatura ambiente ($23\text{ }^{\circ}\text{C}$) no período denominado overnight, que equivale a aproximadamente 12 horas no período noturno. Espectros foram obtidos a partir da folha de alumínio contendo a amostra seca, utilizando-se o espectrômetro Cary 630 FTIR (Agilent Technologies, Inc.) equipado com um acessório de amostragem por reflectância total atenuada de diamante (ATR). A faixa espectral foi de 4.000 a 650 cm^{-1} , no modo absorbância com resolução de 4 cm^{-1} , com 32 scans para *background* e amostra. A cada análise, a janela de amostragem de diamante e a ponta de prensa da amostra foram limpas com etanol 70% v/v. O espectro MIR foi medido em triplicata, com um tempo médio de 90 segundos para cada amostra, fornecendo um conjunto de dados com 795 linhas (amostras em triplicata) e 1.798 colunas (variáveis).

4.2.3 Análises multivariadas

O espectro médio de cada triplicata foi obtido e processado para correção da linha de base (airPLS).²² Este procedimento pode reduzir o impacto de artefatos de espalhamento e deslocamentos indesejáveis em conjuntos de dados MIR. Isso é importante para estudos biológicos devido ao comprimento de onda MIR ($2,5\text{-}25\text{ }\mu\text{m}$) incluir dimensões de uma célula biológica, dando condições para dispersão.¹¹

Para análises multivariadas, os espectros foram cortados na região da *biofingerprint* (1.800-900 cm^{-1}), uma vez que esta região contém informações biológicas relevantes.^{6,10,12,17} Com espectros MIR médios e recortados, obtivemos um conjunto de dados com 265 linhas (amostras) e 484 colunas (variáveis). Para pré-processamento os métodos de pré-processamento foram avaliados individualmente ou em combinação com mais de um método: centralização na média, primeira e segunda derivadas,²³ padronização normal de sinal (SNV),²⁴ normalização (pela norma) e correção de dispersão multiplicativa (MSC).²⁵ Todo o processamento foi realizado com a versão Matlab13A (The MathWorks, Inc.), com algumas interfaces de ferramentas para modelagens²⁶ e *scripts* elaborados em nosso laboratório. As etapas foram divididas em análise não supervisionada, para identificar tendências no conjunto de dados, e análises supervisionadas para classificar as amostras como positivas ou não positivas para infecções por SARS-CoV-2.

4.2.4 Análise não supervisionada

Devido à complexidade das informações espectroscópicas em amostras biológicas, escolhemos o algoritmo floresta aleatória como reconhecimento de padrão não supervisionado. Neste trabalho, utilizamos o modelo de floresta aleatória não supervisionada de acordo com a metodologia proposta por Afanador *et al.*²⁷ para visualizar as semelhanças e diferenças entre amostras. Para isso, a matriz original foi concatenada com conjunto de *outliers* sintéticos gerados e modelada por meio do modelo RF com 1000 árvores e 10% das variáveis como parâmetro *mtry*. O modelo foi avaliado e utilizado para calcular a matriz de proximidade e dissimilaridade. Por fim, as tendências dos dados foram avaliadas por meio da matriz de dissimilaridade em espaços reduzidos pela análise dos gráficos de coordenadas da PCoA com a distância euclidiana das amostras.

A partir do modelo URF foi realizada a identificação das variáveis de maior relevância pelo método de seleção de variáveis proposto por Lovatti e colaboradores.²⁸ Para cada árvore construída como um classificador independente, no modelo URF classificando entre amostras originais e sintéticas, foram identificadas as variáveis que foram selecionadas pelo índice de Gini na etapa de crescimento da árvore. Cada vez que uma variável foi identificada como nó preditor

de uma árvore, foi adicionado um ponto de frequência para esta variável. Ao final, a frequência total de cada variável foi normalizada pelo número de árvores e a frequência relativa foi calculada. Para extrair dessa informação um número de variáveis de maior relevância adotou-se um ponto de corte com base na frequência relativa. Variáveis com frequência relativa acima de 50% foram identificadas, uma vez que passaram pelo processo de seleção aleatória, seguida de seleção pelo índice de Gini em um número significativo de classificadores independentes.

4.2.5 Análise supervisionada

Para análise supervisionada, utilizamos os resultados do RT-qPCR para atribuição de classe das amostras. Para isso, calculamos a média do *Cycle Threshold* (CT) dos genes-alvo da análise RT-qPCR (Gene N e Gene ORF1ab). Para amostras com CT média < 37 foram atribuídas classe um "Positivo", e para amostras com CT média > 37 foram atribuídas classe dois "Negativo". Em seguida, o conjunto original de dados e vetor das classes foram divididos em um conjunto de treinamento (70%) e um conjunto de testes (30%) pelo método *Kennard Stone*²⁹, mantendo a proporção original entre as duas classes. Em seguida, o conjunto de dados de treinamento e teste foram pré-processados, e *outliers* identificados pelo gráfico de controle do Q *residual* e T^2 Hotelling. Foram realizados métodos de seleção de variável (GA, SPA e otimização de enxame de partículas (PSO, do inglês *particle swarm optimization*) associados aos modelos de classificação (PLS-DA e LDA).

4.2.6 Métodos de seleção de variáveis

Os métodos de seleção de variáveis algoritmo genético (GA) e algoritmo de projeção sucessiva (SPA) foram associados à análise discriminante linear (LDA) para constituir os modelos de classificação GA-LDA e SPA-LDA. No modelo GA, as variáveis são codificadas aleatoriamente binariamente com 0 e 1. A identificação das variáveis mais relevantes é baseada na seleção natural darwiniana, na qual as variáveis selecionadas são aquelas que possuem o melhor código genético (variáveis codificadas com 1) por meio da evolução dos indivíduos ao longo de muitas gerações. Um novo conjunto, menor e mais correlacionado com o parâmetro de interesse será utilizado para análise discriminante pelo LDA.^{30,31} A seleção das variáveis por GA foi realizada com 100 gerações e cada uma contendo 200

cromossomos, *crossover* e probabilidades de mutação foram fixadas em 60 e 10%, respectivamente. Uma seleção foi escolhida após a realização de três ciclos de seleção, devido à aleatoriedade do método.

O algoritmo SPA tem como objetivo eliminar variáveis cujas informações são redundantes, ou seja, informações correlacionadas, reduzindo a colinearidade dos dados. Começa com uma variável que incorpora outra variável, até que o número desejado de variáveis seja atingido. Cadeias de variáveis são geradas e avaliadas por meio de uma equação de risco (Equações 01 e 02). O novo conjunto com o menor valor será utilizado para a construção do modelo SPA-LDA.^{14,32-34}

$$G = \frac{1}{K_v} \sum_{K=1}^{K_v} gK \quad (\text{Eq. 01})$$

onde gK é definido como

$$gK = [r^2(x_k, \mu_{Ik})] / [\min_{Ij \neq I} r^2(X_k, \mu_{Ij})] \quad (\text{Eq. 02})$$

A otimização de enxame de partículas é outra técnica que usa codificação binária para selecionar variáveis. O PSO simula no conjunto de variáveis um comportamento observado em animais em busca de alimento ou abrigo, para encontrar a melhor superfície de resposta adequada para o problema de otimização. Os indivíduos "voam" no espaço de pesquisa, se comunicam uns com os outros e atualizam seus X_i posições, e v_i velocidades, até que a solução ótima seja encontrada, ou a condição de parada seja alcançada (Equações 03 e 04). As atualizações ocorrem por meio de fatores de inércia w , o fator da memória individual c_1 (fator cognitivo), um conhecimento social c_2 (fator social). Estes dois últimos fatores $r1$ e social $r2$ pesos, respectivamente, e eles são criados aleatoriamente com valores entre 0 e 1. Além disso, as melhores posições de cada partícula são dadas pelo **pbest** vetor, que serve para atualizar a melhor posição da partícula (aptidão) individualmente. Finalmente, pelo valor de aptidão, o enxame atualiza sua posição (**Gbest**^{it}).³⁵

$$X_i^{it+1} = X_i^{it} + v_i^{it+1} \quad (\text{Eq. 03})$$

$$v_i^{it+1} = w * v_i^{it} + (c_1^{it} * r1^{it} * (pbest_i^{it} - X_i^{it})) + (c_2^{it} * r2^{it} * (Gbest^{it} - X_i^{it})) \quad (\text{Eq. 04})$$

Para a seleção de variáveis, um vetor de resposta deve ser usado, e a seleção pode ser executada uma ou várias vezes, reduzindo o conjunto de variáveis selecionadas. O número de partículas (*popsiz*e) e as iterações também são otimizados. Finalmente, a combinação do PSO e do modelo de análise discriminante de mínimos quadrados parciais (PSO-PLS-DA) é adquirida por um conjunto de variáveis selecionadas.

Para a seleção de variáveis pelo PSO, utilizamos como vetor de resposta a média CT e PSO foi executado com cinco ciclos de seleção, utilizando método de autoescalamento, número de partículas (*popsiz*e) e iterações igual a 10.

4.2.7 Modelos de classificação linear

PLS-DA é um método de classificação linear supervisionada baseado na regressão PLS, onde as informações de cada classe, ou seja, o vetor y , é utilizado como critério de classificação amostral.^{26,36,37} O PLS-DA utiliza variáveis latentes (VL) para extrair as informações importantes (variância) da matriz de dados (X) combinado com o vetor de resposta (y). No PLS-DA, a probabilidade é calculada para cada classe e a classificação das amostras é realizada pela classe com maior probabilidade. O limite de escolha de classe (pertence (1) ou não pertence (0)) pode ser calculado por Teorema de Bayes, assumindo uma distribuição normal dos dados.^{26,36} LDA é uma metodologia que pode mapear conjuntos de vetores de projeção discriminante de alta dimensão (eixos de projeção) e realocá-los em espaço de baixa dimensão. Os eixos de projeção, por meio dos projetados, formam a dispersão máxima entre as diferentes classes e a dispersão mínima entre elas.³⁷⁻³⁹

Vários modelos foram adquiridos para um conjunto de variáveis selecionadas (SPA-LDA; GA-LDA; PLS-DA; PSO-PLS-DA). Para PLS-DA e PSO-PLS-DA utilizamos validação cruzada pelo k-fold ($k=10$) para otimizar a variável latente (VL) por meio da taxa de erro. As configurações para treinar os modelos são mostradas na Tabela 4.1.

Tabela 4.1 Configurações dos modelos.

Modelos	Configurações					
	Pré-processamento	Class_prob ou método de atribuição	Método	Número de componentes ou variáveis latentes (VL)	Método de validação cruzada	Grupos de validação cruzada
GA-LDA	Nenhum	Probabilidade previa / proporcional	Linear	----	<i>Venetian blinds</i>	10
PLS-DA	Centralização na média	Limiares baseados no Teorema de Bayes		7	<i>Venetian blinds</i>	10
SPA-DA	Nenhum	Probabilidade proporcional	Linear	----	<i>Venetian blinds</i>	10
PSO-PLSDA	Centralização na média	Limiares baseados no Teorema de Bayes		9	<i>Venetian blinds</i>	10

4.2.8 Classe de consenso

Da mesma forma que uma fusão de modelos ou método denominado MPA (do inglês, *model population analysis*),⁴⁰ que é uma combinação de modelos ao nível de decisão, os modelos individuais foram construídos, e suas previsões foram integradas em uma única resposta final.⁴¹ Em seguida, cada amostra foi classificada considerando a classe prevista e sua probabilidade calculada entre todos os modelos (Equação 05). Para isso, avaliou-se uma combinação da decisão individual de três modelos de classificação (GA-LDA, PLS-DA e PSO-PLS-DA) para uma decisão final de classe prevista para as amostras. A classe definida para a amostra de cada modelo construído é mostrada na Tabela S.1 (Capítulo 7, Anexos), em que cada amostra foi disposta para dois resultados (positivo (1) ou negativo (2)).

$$Prob_{mult} = \frac{\sum(Prob_{class})}{\sum(Prob_{clas})} \quad (\text{Eq. 05})$$

Onde *Prob* é a probabilidade calculada do modelo para cada classe. *Prob_{mult}* é o fator de decisão obtido entre a classe 1 e a classe 2, e o inverso. O maior valor de razão fornece classe de decisão final para a amostra.

Os modelos de classificação foram avaliados pelas métricas de exatidão, sensibilidade, especificidade e outras métricas de característica de desempenho. Para avaliar a significância estatística dos modelos foi aplicado o teste de

permutação de y . Para isso, o vetor de classe do conjunto de dados de treinamento foi permutado, um modelo foi construído e a classe prevista foi fornecida por meio do modelo permutado. Com a permutação se perde a relação entre a informação química refletida no espectro e a classe atribuída à respectiva amostra. Os parâmetros de desempenho do modelo original são esperados fora da distribuição dos modelos permutados. Este teste de permutação de y foi avaliado pela métrica F1 score para a classe 1 (positivo). Essa métrica foi usada para representar estatisticamente o conjunto de parâmetros de desempenho por meio de um único valor escalar.

Por fim, os modelos avaliados foram utilizados para estimar o diagnóstico em um novo conjunto de dados. Para isso, os modelos foram aplicados a um conjunto de dados de 59 amostras recém-coletadas aleatoriamente (177 triplicatas espectrais). O novo espectro foi processado como o conjunto de dados de treinamento e validação e foi classificado por modelos individuais e de consenso. O RT-qPCR do novo conjunto de dados foi obtido, e as métricas para previsão calculadas.

4.3. Resultados e discussão

Neste estudo, identificamos com sucesso uma estrutura nos dados pelo modelo URF. Em seguida, construímos modelos de classificação linear e aplicamos para diagnosticar amostras de saliva entre positivo e negativo para infecções COVID-19 por meio de espectros MIR. Os procedimentos para identificar e remover *outliers* (antes e depois dos dados de pré-processamento) resultaram em um conjunto de dados com 237 amostras. A Tabela 4.2 descreve perfis amostrais com grupos por sexo e comorbidades, com números e percentuais em seu respectivo subconjunto. Há 138 amostras de pacientes com diagnóstico positivo e 99 com diagnósticos negativos de RT-qPCR para COVID-19. Além disso, das amostras positivas, 67% são mulheres e 33% homens. Enquanto das amostras negativas 71% são mulheres e 29% homens. Entretanto, os dados não apresentaram evidência estatística suficiente ($\alpha=0,05$) para rejeitar a hipótese nula de que a distribuição dos diagnósticos é independente do sexo do paciente por meio do teste χ^2 (p-valor de 0,5095).

Tabela 4.2 Dados gênero e comorbidades do coorte do estudo.

		Número (%)		
		Total (n=237)	Positivo	Negativo
Gênero	Feminino	162 (68%)	92 (57%)	70 (43%)
	Masculino	75 (32%)	46 (61%)	29 (39%)
Comorbidades	Hipertensão	58 (24%)	33 (57%)	25 (43%)
	Diabetes	20 (8%)	13 (65%)	7 (35%)
	DPOC	15 (6%)	10 (67%)	5 (33%)
	Obesidade	13 (5%)	8 (62%)	5 (38%)

DPOC = Doença pulmonar obstrutiva crônica

Os participantes deste estudo foram pacientes atendidos pelos serviços de saúde entre 20 e 97 anos. No entanto, a maior frequência é entre pessoas de 30 a 60 anos (Figura S1) e apenas um paciente tinha mais de 90 anos. No teste χ^2 não há evidência estatística suficiente ($\alpha=0,05$) para rejeitar a hipótese nula de que a distribuição dos diagnósticos é independente da idade do paciente (p-valor 0,5541). Os valores de CT dos genes-alvo são distribuídos entre 12,33 e 40,38 de Gene N e 10,99 a 41,56 para o gene ORF1ab, além dos valores não detectados que foram atribuídos como 42 (Figura S2).

O perfil de espectro médio (normalizado) dos conjuntos de dados, na região de *biofingerprint* (Figura 4.2) e regiões espectrais completas (Figura S3) apresentam as bandas características de absorção de grupos funcionais que caracterizam biomoléculas presentes na saliva. É possível observar pelos espectros médios algumas regiões importantes com diferença na intensidade dos sinais, como a região de Amida I ($\sim 1650 \text{ cm}^{-1}$). Para este estudo, espera-se alterações principalmente nas regiões espectrais de bandas características de proteínas e de lipídios, pelas características do vírus SARS-Cov-2.⁴

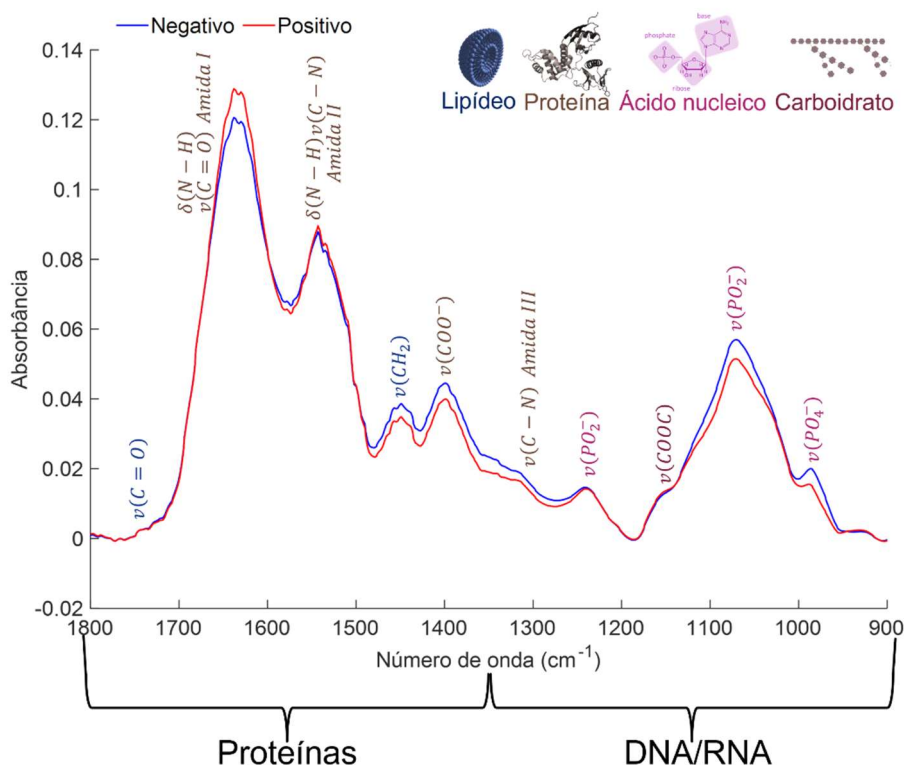


Figura 4.2 Espectros MIR médios normalizados do conjunto de dados de saliva (n=237) de pacientes do coorte deste estudo (linha azul: espectro médio do conjunto de amostras negativas para COVID-19, n=99); linha vermelha: espectro médio do conjunto de amostras positivas COVID-19, n=138).

Esta região de *biofingerprint* é importante para estudos biológicos devido a informações sobre vibração molecular de biomarcadores, como lipídios ($\sim 1.750 \text{ cm}^{-1}$), carboidratos ($\sim 1.155 \text{ cm}^{-1}$), proteínas (Amida I, $\sim 1.650 \text{ cm}^{-1}$, Amida II, $\sim 1.550 \text{ cm}^{-1}$, Amida III, $\sim 1.260 \text{ cm}^{-1}$), além de DNA/RNA (~ 1.225 e $\sim 1.080 \text{ cm}^{-1}$).^{6,8} Principais atribuições de bandas são descritas na Tabela 4.3 para os dados estabelecidos neste estudo nesta região de *biofingerprint*, além de outras bandas presentes nesses espectros. O aspecto dos espectros originais e após a correção de linha de base são demonstrados na Figura S-4.

Tabela 4.3 Principais atribuições de bandas presentes nos espectros.^{6,8}

Banda	Atribuições ^{6,8}
$\sim 3.275 \text{ cm}^{-1}$	Alongamento simétrico O–H
$\sim 3.200 - 3.550 \text{ cm}^{-1}$	Vibrações simétricas e assimétricas atribuídas à água
$\sim 2.930 \text{ cm}^{-1}$	Alongamento C–H
$\sim 2.800 - 3.000 \text{ cm}^{-1}$	C–H de lipídios
$\sim 2.100 \text{ cm}^{-1}$	Uma combinação de rotação e dobramento de O–H (água)
$\sim 1.750 \text{ cm}^{-1}$	Lipídios: $\nu(\text{C}=\text{C})$
$\sim 1.650 \text{ cm}^{-1}$	Amida I: $\nu(\text{C}=\text{O})$
$\sim 1.550 \text{ cm}^{-1}$	Amida II: $\delta(\text{N–H})$ combinado com $\nu(\text{C–N})$
$\sim 1.450 \text{ cm}^{-1}$	Grupo metil em proteína: $\delta[(\text{CH}_3)]$ assimétrico
$\sim 1.400 \text{ cm}^{-1}$	Grupo metil em proteína: $\delta[(\text{CH}_3)]$ simétrico
$\sim 1.250 - 1.260 \text{ cm}^{-1}$	Amida III: $\nu(\text{C–N})$
$\sim 1.155 \text{ cm}^{-1}$	Carboidratos: $\nu(\text{C–O})$
$\sim 1.225 \text{ cm}^{-1}$	DNA e RNA: $\nu_{as}(\text{PO}_2^-)$
$\sim 1.080 \text{ cm}^{-1}$	DNA e RNA: $\nu_s(\text{PO}_2^-)$
$\sim 1.030 \text{ cm}^{-1}$	Glicogênio: $\nu_s(\text{C–O})$
$\sim 971 \text{ cm}^{-1}$	Ácidos nucleicos e proteínas: $\nu_s(\text{PO}_4)$
$\sim 960-966 \text{ cm}^{-1}$	C–O, C–C, desoxirribose

ν_s =alongamento simétrico; ν_{as} = alongamento assimétrico; δ = deformação angular no plano.

4.3.1 Floresta aleatória não supervisionada

O modelo URF foi aplicado para identificar uma possível estrutura do conjunto de dados de espectros. Uma vez que os dados foram modelados com *bootstrapping* de amostras e seleção aleatória de variáveis na presença de *outliers* sintéticos, esta estrutura permite uma distinção entre esses dados originais e sintéticos. O modelo RF distinguiu conjunto de dados originais e sintéticos com exatidão de 98,2%, sensibilidade de 98,9%, e especificidade de 97,5%, indicando a presença de estrutura nos dados.

A partir desse modelo URF, a matriz de proximidade permitiu a análise de PCoA (80% de variância), e as amostras foram projetadas em três dimensões pelo gráfico de escores PCoA (Figura 4.3). É possível visualizar diferentes grupos. No

entanto, não há uma distância satisfatória entre as classes em PCo 1, que seria necessário para uma distinção eficiente entre estes grupos.

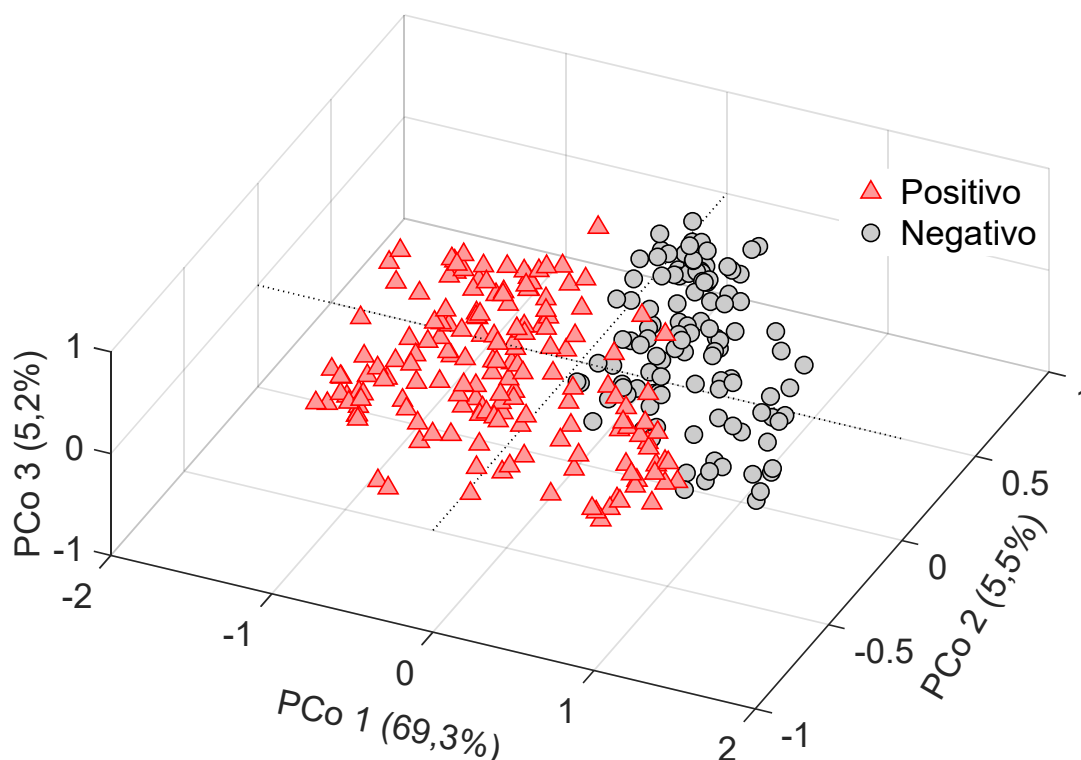


Figura 4.3 Análise de coordenadas principais (PCoA) a partir de modelo URF de conjunto de dados espectros MIR de saliva (n=246).

Neste modelo URF, foram identificadas 82 variáveis com maior frequência (Figura S-5). Dessas variáveis há bandas características, como algumas associadas às regiões lipídicas (1.785 a 1.729 cm^{-1} ; alongamento C=C e C=O dos grupos éster), proteínas (1.680 e 1.718 a 1.705 cm^{-1} : alongamento C=O e C-N; 1.600 a 1.250 cm^{-1} : amida I, II e III). Além disso, as bandas características de ácido nucleico (1.612 - 1.606 cm^{-1} : vibração da adenina no DNA, 1.244 - 1.100 cm^{-1} : alongamento PO_4 de grupos de fosfodiéster; 1.025 a 1.021 cm^{-1} : alongamento C-O (carboidratos); 961 cm^{-1} : desoxirribose, e 930 a 909 cm^{-1} : bandas de alongamento fosfodiéster).^{6,8} Vale ressaltar que há uma limitação na interpretação dessas variáveis a partir de um modelo URF, uma vez que o modelo foi treinado para promover distinção entre todo o conjunto de amostras originais, sem distinção entre positivas e negativas, e um conjunto aleatório sintético.

4.3.2 Análises supervisionadas

Aplicamos modelos lineares para classificar amostras com métodos de seleção de variáveis (SPA-LDA, GA-LDA), redução de dimensão (PLS-DA) e uma combinação de seleção variável e redução de dimensões (PSO-PLS-DA). SPA-LDA, GA-LDA e PLS-DA são os métodos de classificação mais aplicados em estudos biológicos.⁶ O mesmo conjunto de treinamento e de teste foi utilizado para cada modelo. Vários métodos de pré-processamento foram testados, mas a suavização pela segunda derivada (21 pontos de janela e polinômio de segundo grau), cujo aspecto está demonstrado na Figura S-6, apresentou os melhores resultados nos modelos de classificação. A partir da resposta de modelos individuais, a classe de consenso foi atribuída às amostras por meio da probabilidade de modelos. Do conjunto de treinamento, 35 amostras (21%) foram classificadas erradas, caracterizando falso positivo e falso negativo, e no conjunto de testes, esse número diminuiu em 12 (17%) de acordo com a matriz de confusão de classe de consenso (Tabela 4.4) e Figura S-7. A matriz de confusão dos modelos individuais é apresentada na Tabela S-2.

Tabela 4.4 Matriz de confusão de modelos de classificação de consenso.

Classe real	VP	VN	FP	FN
Treinamento				
Positivo	82	48	20	15
Negativo	48	82	15	20
Teste				
Positivo	38	22	9	3
Negativo	22	38	3	9

VP = Verdadeiro positivo; VN = Verdadeiro negativo; FP = Falso positivo; FN = Falso negativo.

As principais métricas de características de desempenho dos modelos e parâmetros individuais da classificação de consenso são mostradas na Tabela 4.5 para o conjunto de treinamento e teste. Entre os modelos individuais, o GA-LDA e o PSO-LDA destacam-se com parâmetros de maior qualidade. O coeficiente de correlação de Matthews (MCC) tem sido usado principalmente para casos com número desbalanceado de amostras entre as classes.^{42–44} Este parâmetro usa a matriz de confusão para calcular uma correlação entre classes reais e estimadas.

Um valor MCC próximo a zero sugere que a previsão não foi melhor do que uma previsão aleatória.⁴²⁻⁴⁴ Os modelos GA-LDA, PSO-PLS-DA e SPA-LDA apresentaram valor MCC > 0,5, além do parâmetro obtido da classe de consenso para decisão final.

Tabela 4.5 Parâmetros de avaliação de desempenho de modelos de classificação de conjuntos de treinamento (linha sem sombreamento) e teste (linhas sombreadas).

Modelo	Amostras			Parâmetros					
	POS	NEG	outlier	SENS %	ESPEC %	PREC CL.1 %	PREC CL.2 %	EXAT %	MCC
SPA-LDA	97	68	7	69,1	63,2	72,8	58,9	66,7	0,320
	41	31	2	87,8	67,7	78,3	80,8	79,8	0,572
GA-LDA	97	68	7	86,6	67,6	79,2	77,9	78,8	0,557
	41	31	2	95,1	70,9	81,2	91,7	84,7	0,694
PLS-DA (7 VL)	97	68	7	70,0	76,5	80,9	64,2	72,7	0,459
	41	31	2	75,6	74,2	79,5	69,7	75,0	0,495
PSO PLS-DA (9 VL)	97	68	7	79,4	76,5	82,8	72,2	78,2	0,554
	41	31	2	82,9	74,2	80,9	76,7	79,2	0,574
Classe de consenso	97	68	7	82,5	75,0	82,0	75,0	79,0	0,575
	41	31	2	93,0	74,0	83,0	88,0	85,0	0,689

SENS = Sensibilidade; ESPEC = Especificidade; PREC CL.1 = Precisão de Classe 1, PREC CL.2 = Precisão de classe 2; EXAT = Exatidão; MCC = coeficiente de correlação de Mathews; VL = Variáveis latentes.

Bandas distintas foram selecionadas pelos modelos. No SPA foram selecionadas cinco variáveis, na GA 34, e 45 pelo PSO. Nos modelos PLS-DA foram identificadas variáveis mais importantes por meio dos valores de coeficiente. Esta identificação foi realizada separadamente para as classes 1 e classe 2. Do modelo PLS-DA sem seleção de variável, foram destacados 63 números de onda relacionados à classe 1 e 37 para a classe 2. Enquanto no PLS-DA após o método PSO foram destacadas 6 variáveis relacionadas à classe 1 e 5 à classe 2. As regiões lipídicas são destacadas nessas seleções (1.707-1.792 cm⁻¹), principalmente em GA (Figura 4.4 A), PSO (Figura 4.4 C), e PLS-DA (Figura 4.4 D – E). A partir do PLS-DA (Figura 4.4 D) esta região é mais importante para distinguir classe 1, ou seja, classe

positiva. Há poucos estudos sobre evidências de uma relação entre os níveis de triglicerídeos e infecções por COVID-19 em biofluidos ou outras fontes.^{45,46} A região de amida I ($\sim 1.650\text{ cm}^{-1}$) foi selecionada pelos métodos GA (Figura 4.4 A) e PSO (Figura 4.4 E). Além disso, essa região é destacada no PLS-DA e PSO-PLS-DA (Figura 4.4 E) para a classe 2, ou seja, negativa. As regiões que apresentam maiores valores de coeficiente PLS-DA (Figura 4.4 D) para a classe 1 são bandas mais próximas de 1.400 cm^{-1} , ou seja, região proteica, e mais próximas de 1.200 cm^{-1} , 1.155 cm^{-1} e 950 cm^{-1} que compreendem carboidratos, DNA/RNA e ácido nucleico, respectivamente. Além disso, o PSO-PLS-DA reduziu 90% do número de variáveis mais importantes para a classe 1 e 86,5% para a classe 2, quando comparado ao PLS-DA sem seleção variável.

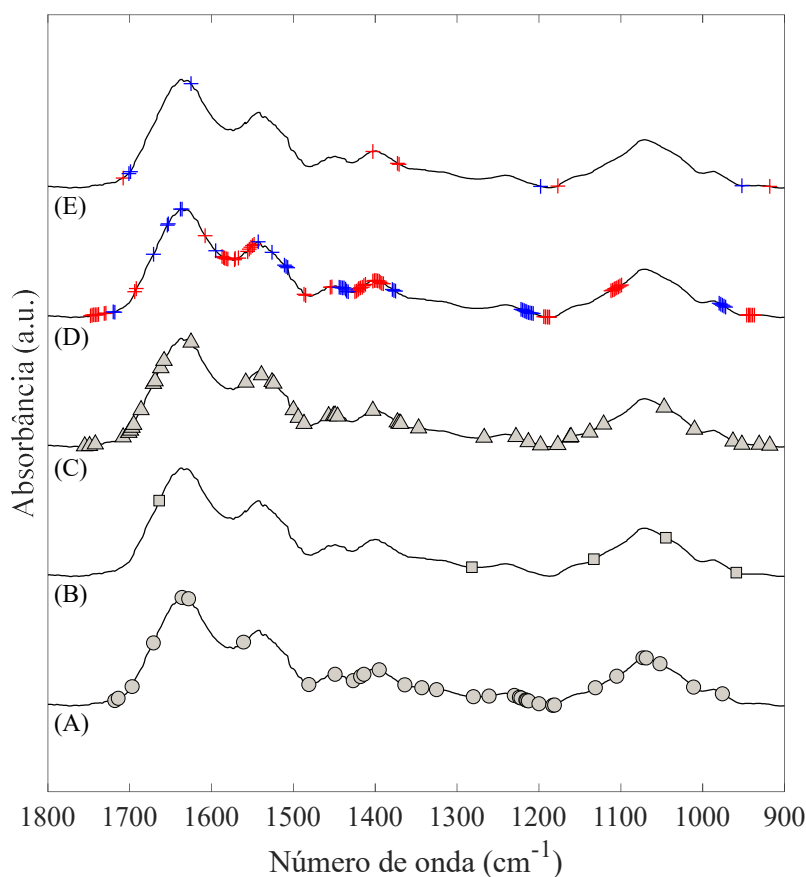


Figura 4.4. Variáveis mais importantes (coeficiente PLS-DA) e selecionadas por meio de métodos de seleção de variáveis e modelos de classificação. (A) Variáveis selecionadas pelo método GA-LDA; (B) variáveis selecionadas pelo método SPA-LDA; (C) Variáveis selecionadas pelo método PSO; (D) Variáveis mais importante para classe 1 (+ vermelha) e classe 2 (+ azul) por meio de análise dos valores de coeficientes PLS-DA; (E) Variáveis mais importante para classe 1 (+ vermelha) e classe 2 (+ azul) por meio da análise dos valores de coeficientes PSO-PLS-DA.

Algumas das variáveis selecionadas correspondem às descritas em Baraúna *et al.*²¹ ($\sim 1.429\text{ cm}^{-1}$, $\sim 1.220\text{ cm}^{-1}$, $\sim 1.069\text{ cm}^{-1}$). No entanto, Baraúna *et al.*²¹ utilizaram espectros de *swab* na região da boca, com algumas bandas mais bem definidas na região de $1.100\text{--}900\text{ cm}^{-1}$. Considerando que um esfregaço celular possa apresentar maior concentração de componentes, isso pode explicar uma diferença espectral e uma possível maior variância nesta região. Além disso, Wylie *et al.*⁷ relataram cópias de RNA SARS-CoV-2 mais altas na saliva em comparação com *swab* nasofaríngeos. Este vírus tem um tropismo preferencial para células epiteliais das vias aéreas humanas, e as glândulas salivares podem ser um alvo potencial para SARS-CoV-2.^{2,3,5,7,47}

Em outro artigo com classificação de amostras biológicas para o diagnóstico de COVID-19 com uso de MIR, Zhang *et al.*⁴⁸ obtiveram distinção entre espectros MIR de amostras de soro sanguíneo por meio do modelo PLS-DA (sensibilidade de 83% e especificidade de 98%) com dados processados pela segunda derivada entre pacientes do grupo controle (pessoas saudáveis) e pacientes com diagnóstico confirmado de COVID e sintomas respiratórios. A região mais importante para os modelos foi de $1.450\text{--}1.650\text{ cm}^{-1}$ e $1.050\text{--}1.100\text{ cm}^{-1}$. Entretanto, além das amostras invasivas e maior tempo de análise, os autores enfatizam neste trabalho que espectros de pacientes assintomáticos ou diagnosticados com intervalo precoce de dias de sintomas podem não ser corretamente identificados pelo modelo e que isso depende de uma decisão do limite do modelo PLS-DA. Esse desafio está de acordo com nossos resultados, pois mesmo com exatidão aceitável, os modelos apresentam uma alta taxa de falso-positivo.

Neste estudo, considerando que todos os participantes apresentavam sintomas de infecção respiratória (Tabela S-3), espera-se que a distinção promovida pelo método proposto seja em função das mudanças bioquímicas relevantes provocadas pela presença do vírus SARS-CoV-2 no sistema biológico. Ademais, a prevalência de valor negativo, ou a precisão para a classe 2 (negativo) foi 88%, ou seja, das amostras classificadas pelo modelo como negativo, 88% são verdadeiramente negativas. Isso sugere que apesar de apresentarem sintomas semelhantes, o modelo distinguiu amostras negativas para SAR-CoV-2 com boa precisão. Neste caso, uma alta taxa de falso negativo pode ser um problema de maior preocupação, uma vez que uma pessoa infectada classificada como saudável,

pode potencializar a disseminação do vírus. Por essas razões, parâmetros como sensibilidade (93%) e prevalência de valor negativo (88%) são potenciais indicadores de que os biomarcadores modelados no espectro MIR são relacionados à presença do vírus SARS-CoV-2 nas amostras de saliva.

O coorte de pacientes apresentou variabilidade de sintomas entre médio a moderado e dias de sintomas variando entre um e dez. Entretanto, há uma maior concentração entre três e seis dias, e alguns *outliers* com mais de dez dias de sintomas (Figura S-8). Em amostras com três dias de sintomas há um número mais alto de falso-negativo. Entre quatro e cinco dias há um maior número de falso positivo, e de seis dias de sintomas observa-se uma tendência para um aumento de falso negativo (Figura 4.5). No entanto, o teste $\chi^2(\alpha=0,05)$ mostra que não há evidência estatística suficiente para descartar a hipótese nula de que a distribuição de amostras classificadas erradas por meio da classe de consenso é independente dos dias de sintomas (p-valor = 0,4224).

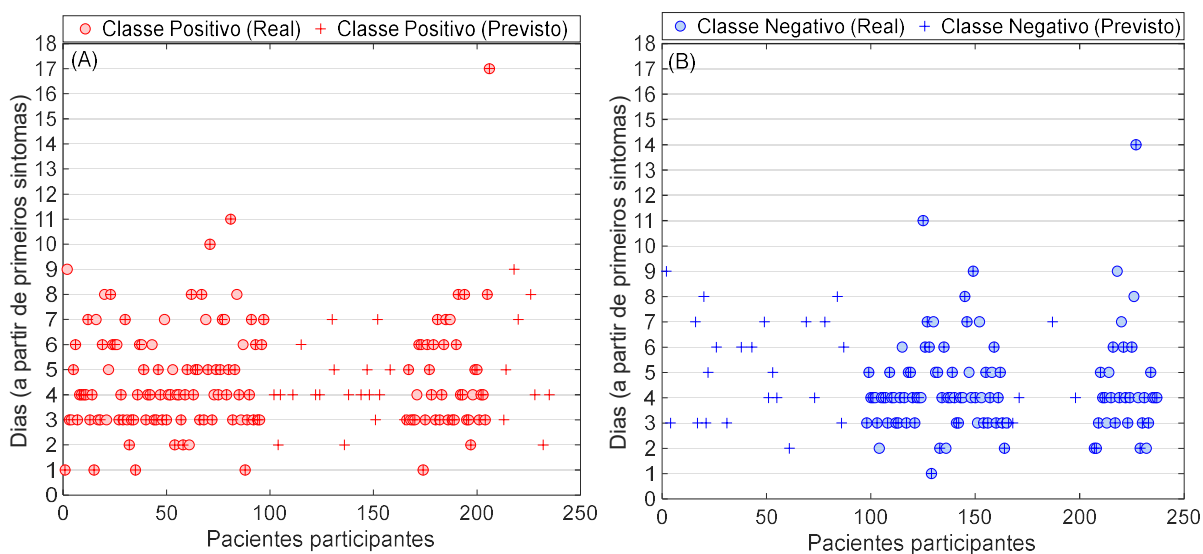


Figura 4.5 Gráfico das classes real e prevista e dias dos primeiros sintomas do conjunto de dados ($n=237$). (A) Marcadores de classe positivos; (B) Marcadores de classe negativos.

Para avaliar uma aplicação clínica, os modelos foram testados em um novo conjunto de dados ($n=59$, de pacientes sintomáticos de mesma região e serviços de saúde que os anteriores), para classificação com modelos de classificação individual e decisão final pela classe de consenso. Alguns *outliers* neste novo conjunto de

dados foram identificados e não considerados nesta aplicação ($n=8$, $\sim 13\%$). Após diagnósticos clínicos dessas amostras, calculamos os parâmetros de qualidade desta nova previsão (Tabela S-4). A exatidão foi reduzida em 59% a partir da decisão final e para 63% do modelo PSO-PLS-DA. Isso sugere que os modelos precisam melhorar a robustez. O GA-LDA apresentou TFP mais alto nesta nova aplicação (68%), enquanto outros modelos apresentaram $\sim 50\%$. O PSO-PLS-DA apresentou parâmetros de melhor qualidade nesta nova previsão quando comparado com outros modelos. Semelhantemente, Wood *et al.*⁴ concluíram necessitar de um maior grupo de pacientes para melhorar a sensibilidade e a especificidade do método proposto, mesmo atingindo excelentes parâmetros de modelagem (93% sensibilidade), devido ao pequeno conjunto de dados aplicado (29 amostras positivas de saliva para infecção por SARS-CoV-2 e 28 negativos). Avaliamos neste trabalho um número maior de amostras ($n=237$) e testamos um novo conjunto de dados ($n=59$). Além disso, identificamos regiões espectrais importantes por meio de diferentes métodos de seleção de variáveis, que podem esclarecer a relação existente entre as informações espectrais e a resposta biológica à infecção COVID-19. Ademais, por meio do teste de permutação de y (Figura 4.6), observamos que a classe de consenso contribui para tornar a classificação estatisticamente significativa, comparada ao modelo individual.

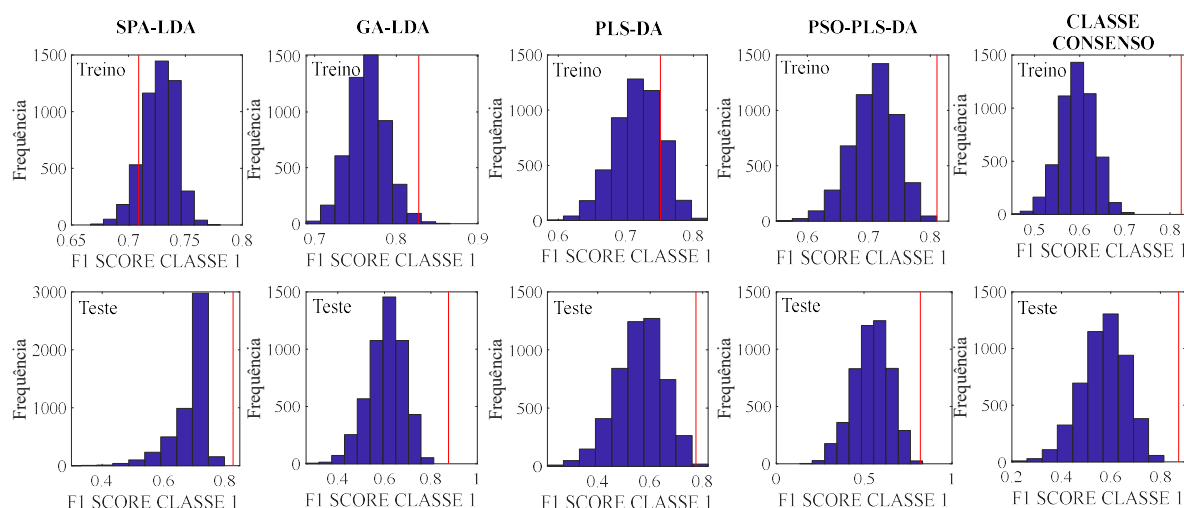


Figura 4.6. Distribuição de parâmetro F1 Score de classe 1 (Positivo) de todos os modelos construídos no teste de permutação de y (histogramas), em comparação com F1 Score de modelos construídos neste estudo (linha vermelha na vertical), para conjunto de treinamento e de teste.

4.4. Conclusões

Os métodos de seleção variável e modelos de classificação linear podem identificar amostras positivas de saliva com exatidão de 83% e precisão de 80 e 88 % para positivo e negativo para infecção COVID-19, respectivamente. Embora o modelo GA-LDA individual apresente um bom desempenho no conjunto de validação, com 95% de sensibilidade e 81% de precisão, a classe de consenso adiciona robustez para prever novas amostras, uma vez que o GA-LDA apresentou uma taxa de falso-positivo maior (68%). O desempenho dos modelos para estimar classes para um conjunto novo e aleatório de amostras (n=59) não foi equivalente aos obtidos para o conjunto de validação. No entanto, o PSO-PLS-DA apresentou melhor desempenho para estimar as classes (sensibilidade de 77%, especificidade de 48%, e exatidão de 63%). Isso sugere que o PSO-PLS-DA pode ser suficiente como um método de classificação alternativa para a triagem de amostras suspeitas. Seu desempenho no conjunto de validação também sugere isso (sensibilidade 83%, especificidade 74%, exatidão 79%, e valor MCC 0,57).

A sensibilidade da espectroscopia MIR para esta análise tem sido confirmada em estudos recentes com fluidos biológicos. Análise não supervisionada por meio do método URF demonstrou que há uma estrutura específica nos dados espectroscópicos MIR nestas amostras de saliva. Além disso, as análises supervisionadas destacaram regiões espectrais relevantes que estão relacionadas a biomarcadores do vírus e respostas biológicas da infecção.

Uma ampla implementação dessa metodologia requer a identificação de vários fatores confundidores, como os fatores que carregam falsos negativos ou falsos positivos no exame utilizado como referência, a resposta biológica à infecção pela COVID-19, outros tipos de infecções, ou outros vírus, além de pessoas assintomáticas. Nossos resultados mostram que essa metodologia é uma ferramenta potencial para isolar possíveis propagadores da infecção. Isso se deve à possibilidade de diagnóstico em pouco tempo e à redução da demanda por suprimentos. Além disso, a coleta de amostras de saliva pelos próprios pacientes evita a interação direta entre profissionais de saúde e pacientes e pode ser uma alternativa ao rastreamento de pessoas infectadas.

Agradecimentos

Todos os autores gostariam de agradecer aos pacientes que, mesmo em um momento delicado, concordaram voluntariamente em participar deste estudo; aos serviços de saúde e ao LACEN pela análise; e, ao LabPetro (UFES, Brasil) pela realização de medições FTIR (Acordos de Cooperação Técnica nº 0050.0022844.06.4). Os autores também agradecem a todos os hospitais (Hospital Vila Velha, Hospital Universitário Cassiano Antonio Moraes, Pronto Atendimento da Glória, Hospital Roberto Arnizaut Silveiras), funcionários, enfermeiros, médicos e pacientes que aceitaram participar do estudo. Os autores agradecem à Prefeitura de Vila Velha e à Secretaria Municipal de Saúde. Os autores agradecem ao Instituto Capixaba de Ensino, Pesquisa e Inovação em Saúde (ICEPi) e à Secretaria Estadual de Saúde do Espírito Santo (SESA) pelos dados clínicos dos pacientes e resultados do RT-qPCR. Este estudo foi apoiado pela FAPES (151/2020, 165/2021, 1036/2022 e 691/2022) e CNPq (401870/2020-0, 310057/2020-5 e 409700/2022-3).

Material suplementar

No material suplementar (Capítulo 7, item 7.2) são encontrados: várias tabelas nas quais são apresentadas: classe prevista pelos modelos; matriz de confusão dos modelos de classificação; tabela cruzada com teste $\chi^2(\alpha=0,05)$ de sintomas e outros dados do coorte; figuras nas quais não apresentados: histograma da distribuição de idade do coorte; histograma com valor quantitativo dos gene alvo de análise RT-qPCR; espectros com principais bandas em amostras biológicas; aspecto dos espectros após pré-processamento; análise de frequência de variáveis no modelo URF; classe real e classe prevista por amostra; e boxplot com dias de sintomas.

4.5. Referências

- (1) Geneva: World Health Organization. *WHO COVID-19 Explorer*. Available online: <https://worldhealthorg.shinyapps.io/covid/>. <https://worldhealthorg.shinyapps.io/covid/> (accessed 2021-06-30).
- (2) Khiabani, K.; Amirzade-Irani, M. H. Are Saliva and Deep Throat Sputum as Reliable as Common Respiratory Specimens for SARS-CoV-2 Detection? A Systematic Review and Meta-Analysis. *Amer. J. Infect. Control* **2021**, *49*, 1165–1176. <https://doi.org/10.1016/j.ajic.2021.03.008>.
- (3) Cañete, M. G.; Valenzuela, I. M.; Garcés, P. C.; Massó, I. C.; González, M. J.; Providell, S. G. Saliva Sample for the Massive Screening of SARS-CoV-2 Infection: A Systematic Review. *Oral Surg. Oral Med. Oral Pathol. Oral Radiol.* **2021**, *131* (5), 540–548. <https://doi.org/10.1016/j.oooo.2021.01.028>.

- (4) Wood, B. R.; Kochan, K.; Bedolla, D. E.; Salazar-Quiroz, N.; Grimley, S. L.; Perez-Guaita, D.; Baker, M. J.; Vongsvivut, J.; Tobin, M. J.; Bambery, K. R.; Christensen, D.; Pasricha, S.; Eden, A. K.; Mclean, A.; Roy, S.; Roberts, J. A.; Druce, J.; Williamson, D. A.; McAuley, J.; Catton, M.; Purcell, D. F. J.; Godfrey, D. I.; Heruad, P. Infrared Based Saliva Screening Test for COVID-19. *Angew. Chem. Int. Ed. Engl.* **2021**, *60* (31), 17102–17107. <https://doi.org/10.1002/anie.202104453>.
- (5) Vogels, C. B. F.; Watkins, A. E.; Harden, C. A.; Brackney, D. E.; Shafer, J.; Wang, J.; Caraballo, C.; Kalinich, C. C.; Ott, I. M.; Fauver, J. R.; Kudo, E.; Lu, P.; Venkataraman, A.; Tokuyama, M.; Moore, A. J.; Muenker, M. C.; Casanovas-Massana, A.; Fournier, J.; Bermejo, S.; Campbell, M.; Datta, R.; Nelson, A.; dela Cruz, C. S.; Ko, A. I.; Iwasaki, A.; Krumholz, H. M.; Matheus, J. D.; Hui, P.; Liu, C.; Farhadian, S. F.; Sikka, R.; Wyllie, A. L.; Grubaugh, N. D. SalivaDirect: A Simplified and Flexible Platform to Enhance SARS-CoV-2 Testing Capacity. *Med* **2021**, *2* (3), 263-280.e6. <https://doi.org/10.1016/j.medj.2020.12.010>.
- (6) Santos, M. C. D.; Morais, C. L. M.; Nascimento, Y. M. Spectroscopy with Computational Analysis in Virological Studies: A Decade (2006e2016). *Trends in Analytical Chemistry* **2017**, *97* (January), 244–256. <https://doi.org/10.1016/j.trac.2017.09.015>.
- (7) Wyllie AL; Fournier J; Casanovas-Massana A; Melissa Campbell; Maria Tokuyama; Pavithra Vijayakumar; Joshua L. Warren; Bertie Geng; M. Catherine Muenker; Adam J. Moore; Chantal B. F. Vogels; Mary E. Petrone; Isabel M. Ott; Peiwen Lu; Arvind Venkataramam; Alice Lu-Culligan; Jonathan Klein; Rebecca Earnest. Saliva or Nasopharyngeal Swab Specimens for Detection of SARS-CoV-2. *New engl. J. Med.* **2020**, *383*, 1283–1286. <https://doi.org/10.1056/NEJMc2016359>.
- (8) Movasaghi, Z.; Rehman, S.; Rehman, I. U. Fourier Transform Infrared (FTIR) Spectroscopy of Biological Tissues. *Appl. Spectrosc. Rev.* **2008**, *43* (2), 134–179. <https://doi.org/10.1080/05704920701829043>.
- (9) Silverstein, R. M.; Webster, F. X.; Kiemle, D. J.; Bryce, D. L. *Spectrometric Identification of Organic Compounds*, 8th ed.; John Wiley & Sons: New York, 2014.
- (10) Morais, C. L. M.; Lima, K. M. G.; Singh, M.; Martin, F. L. Tutorial: Multivariate Classification for Vibrational Spectroscopy in Biological Samples. *Nat. Protoc.* **2020**, *15* (7), 2143–2162. <https://doi.org/10.1038/s41596-020-0322-8>.
- (11) Butler, H. J.; Smith, B. R.; Fritzsche, R.; Radhakrishnan, P.; Palmer, D. S.; Baker, M. J. Optimised Spectral Pre-Processing for Discrimination of Biofluids via ATR-FTIR Spectroscopy. *Analyst* **2018**, *143* (24), 6121–6134. <https://doi.org/10.1039/c8an01384e>.
- (12) Mitchell, A. L.; Gajjar, K. B.; Theophilou, G.; Martin, F. L.; Martin-Hirsch, P. L. Vibrational Spectroscopy of Biofluids for Disease Screening or Diagnosis: Translation from the Laboratory to a Clinical Setting. *J. Biophotonics* **2014**, *7* (3–4), 153–165. <https://doi.org/10.1002/jbio.201400018>.
- (13) Santos, M. C. D.; Nascimento, Y. M.; Araújo, J. M. G.; Lima, K. M. G. ATR-FTIR Spectroscopy Coupled with Multivariate Analysis Techniques for the Identification of DENV-3 in Different Concentrations in Blood and Serum: A New Approach. *RSC Adv.* **2017**, *7* (41), 25640–25649. <https://doi.org/10.1039/c7ra03361c>.

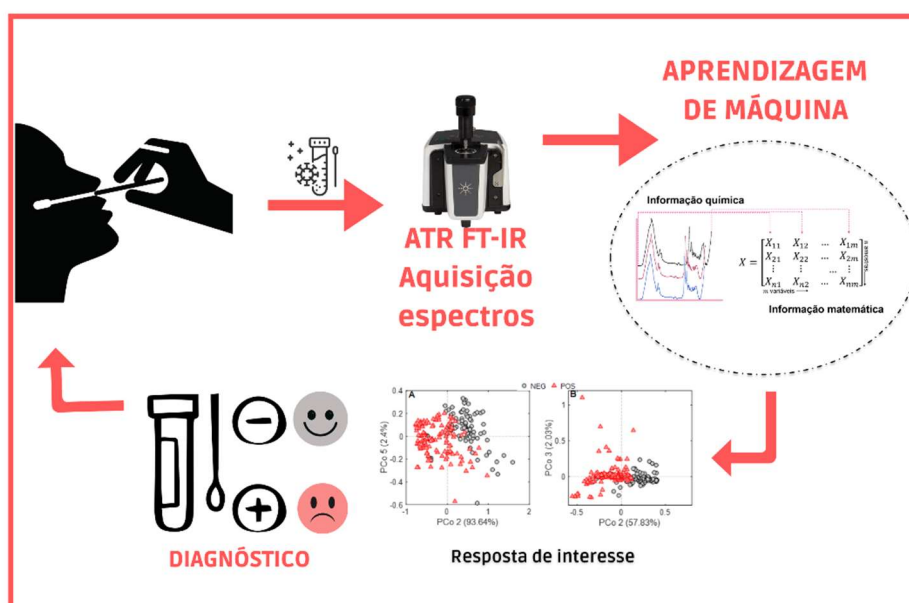
- (14) Marques, A. de S.; Melo, M. C. N. de; Cidral, T. A.; Lima, K. M. G. de. Feature Selection Strategies for Identification of Staphylococcus Aureus Recovered in Blood Cultures Using FT-IR Spectroscopy Successive Projections Algorithm for Variable Selection: A Case Study. *J. Microbiol. Methods*. **2014**, *98* (1), 26–30. <https://doi.org/10.1016/j.mimet.2013.12.015>.
- (15) Huber, M.; Kepesidis, K. v.; Voronina, L.; Božić, M.; Trubetskov, M.; Harbeck, N.; Krausz, F.; Žigman, M. Stability of Person-Specific Blood-Based Infrared Molecular Fingerprints Opens up Prospects for Health Monitoring. *Nat. Commun.* **2021**, *12* (1), 1–10. <https://doi.org/10.1038/s41467-021-21668-5>.
- (16) Sakudo, A.; Suganuma, Y.; Kobayashi, T.; Onodera, T.; Ikuta, K. Near-Infrared Spectroscopy: Promising Diagnostic Tool for Viral Infections. *Biochem. Biophys. Res. Commun.* **2006**, *341* (2), 279–284. <https://doi.org/10.1016/j.bbrc.2005.12.153>.
- (17) Pupeza, I.; Huber, M.; Trubetskov, M.; Schweinberger, W.; Hussain, S. A.; Hofer, C.; Fritsch, K.; Poetzlberger, M.; Vámos, L.; Fill, E.; Amotchkina, T.; Kepesidis, K. v.; Apolonski, A.; Karpowicz, N.; Pervak, V.; Pronin, O.; Fleischmann, F.; Azzeer, A.; Žigman, M.; Krausz, F. Field-Resolved Infrared Spectroscopy of Biological Systems. *Nature* **2020**, *577* (7788), 52–59. <https://doi.org/10.1038/s41586-019-1850-7>.
- (18) Bel'skaya, L. V. Use of IR Spectroscopy in Cancer Diagnosis. A Review. *J Appl Spectrosc* **2019**, *86* (2), 187–205. <https://doi.org/10.1007/s10812-019-00800-w>.
- (19) Peng, X.; Zhou, Y.; Nie, K.; Zhou, F.; Yuan, Y.; Song, J.; Qu, J. Promising Near-Infrared Plasmonic Biosensor Employed for Specific Detection of SARS-CoV-2 and Its Spike Glycoprotein. *New J. Phys.* **2020**, *22* (10). <https://doi.org/10.1088/1367-2630/abbe53>.
- (20) Santos, M. C. D.; Morais, C. L. M.; Lima, K. M. G. ATR-FTIR Spectroscopy for Virus Identification: A Powerful Alternative. *Biomed Spectrosc Imaging* **2020**, *9* (3–4), 103–118. <https://doi.org/10.3233/bsi-200203>.
- (21) Barauna, V. G.; Singh, M. N.; Barbosa, L. L.; Marcarini, W. D.; Vassallo, P. F.; Mill, J. G.; Ribeiro-Rodrigues, R.; Campos, L. C. G.; Warnke, P. H.; Martin, F. L. Ultrarapid On-Site Detection of SARS-CoV-2 Infection Using Simple ATR-FTIR Spectroscopy and an Analysis Algorithm: High Sensitivity and Specificity. *Anal. Chem.* **2021**, *93* (5), 2950–2958. <https://doi.org/10.1021/acs.analchem.0c04608>.
- (22) Zhang, Z. M.; Chen, S.; Liang, Y. Z. Baseline Correction Using Adaptive Iteratively Reweighted Penalized Least Squares. *Analyst* **2010**, *135* (5), 1138–1146. <https://doi.org/10.1039/b922045c>.
- (23) Savitzky, A.; Golay, M. J. E. Smoothing and Differentiation of Data by Simplified Least Squares Procedures. *Anal. Chem.* **1964**, *36* (8), 1627–1639. <https://doi.org/10.1021/ac60214a047>.
- (24) Barnes, R. J.; Dhanoa, M. S.; Lister, S. J. Standard Normal Variate Transformation and De-Trending of near-Infrared Diffuse Reflectance Spectra. *Appl. Spectroscopy* **1989**, *43* (5), 772–777. <https://doi.org/10.1366/0003702894202201>.
- (25) Isaksson, T.; Naes, T. Effect of Multiplicative Scatter Correction (MSC) and Linearity Improvement in NIR Spectroscopy. *Appl. Spectroscopy* **1988**, *42* (7), 1273–1284. <https://doi.org/10.1366/0003702884429869>.

- (26) Ballabio, D.; Consonni, V. Classification Tools in Chemistry. Part 1: Linear Models. PLS-DA. *Anal. Methods* **2013**, *5* (16), 3790–3798. <https://doi.org/10.1039/c3ay40582f>.
- (27) Afanador, N. L.; Smolinska, A.; Tran, T. N.; Blanchet, L. Unsupervised Random Forest: A Tutorial with Case Studies. *J. Chemometrics* **2016**, *30* (5), 232–241. <https://doi.org/10.1002/cem.2790>.
- (28) Lovatti, B. P. O.; Nascimento, M. H. C.; Neto, Á. C.; Castro, E. V. R.; Filgueiras, P. R. Use of Random Forest in the Identification of Important Variables. *Microchemical Journal* **2019**, *145*, 1129–1134. <https://doi.org/10.1016/j.microc.2018.12.028>.
- (29) Kennard, R. W.; Stone, L. A. Computer Aided Design of Experiments. *Tecnometrics* **1969**, *11* (1), 137–148.
- (30) Siqueira, L. F. S.; Araújo Júnior, R. F.; de Araújo, A. A.; Moraes, C. L. M.; Lima, K. M. G. LDA vs. QDA for FT-MIR Prostate Cancer Tissue Classification. *Chemometrics Intell. Lab. Syst.* **2017**, *162*, 123–129. <https://doi.org/10.1016/j.chemolab.2017.01.021>.
- (31) Shaffer, R. E.; Small, G. W. Genetic Algorithms for the Optimization of Piecewise Linear Discriminants. *Chemometrics Intell. Lab. Syst.* **1996**, *35*, 87–104.
- (32) Pontes, M. J. C.; Galvão, R. K. H.; Araújo, M. C. U. The Successive Projections Algorithm for Spectral Variable Selection in Classification Problems. *Chemometrics Intell. Lab. Syst.* **2005**, *78* (1–2), 11–18. <https://doi.org/10.1016/j.chemolab.2004.12.001>.
- (33) Marques, A. S.; Castro, J. N. F.; Costa, F. J. M. D.; Neto, R. M.; Lima, K. M. G. Near-Infrared Spectroscopy and Variable Selection Techniques to Discriminate *Pseudomonas Aeruginosa* Strains in Clinical Samples. *Microchem. J.* **2016**, *124*, 306–310. <https://doi.org/10.1016/j.microc.2015.09.006>.
- (34) Araújo, M. C. U.; Saldanha, teresa C. B.; Galvão, R. K. H.; Yoneyama, T.; Chame, H. C.; Visani, V. The Successive Projections Algorithm for Variable Selection in Spectroscopic Multicomponent Analysis. *Chemometrics Intell. Lab. Syst.* **2001**, *57* (2), 65–73.
- (35) Marini, F.; Walczak, B. Particle Swarm Optimization (PSO). A Tutorial. *Chemometrics Intell. Lab. Syst.* **2015**, *149*, 153–165. <https://doi.org/10.1016/j.chemolab.2015.08.020>.
- (36) Brereton, R. G.; Lloyd, G. R. Partial Least Squares Discriminant Analysis: Taking the Magic Away. *J. Chemometrics* **2014**, *28* (4), 213–225. <https://doi.org/10.1002/cem.2609>.
- (37) Barker, M.; Rayens, W. Partial Least Squares for Discrimination. *J. Chemometrics* **2003**, *17*, 166–173. <https://doi.org/10.1002/cem.785>.
- (38) Chen, L. F.; Liao, H. Y. M.; Ko, M. T.; Lin, J. C.; Yu, G. J. A New LDA-Based Face Recognition System Which Can Solve the Small Sample Size Problem. *Pattern Recognit.* **2000**, *33*, 1713–1726.
- (39) Martõ Ânez, A. M.; Kak, A. C. PCA versus LDA. *IEEE Trans. Pattern Anal. Mach. Intell.* **2001**, *2* (2), 228–233. <https://doi.org/10.1109/34.908974>.
- (40) Deng, B. C.; Yun, Y. H.; Liang, Y. Z. Model Population Analysis in Chemometrics. *Chemometrics and Intelligent Laboratory Systems* **2015**, *149*, 166–176. <https://doi.org/10.1016/j.chemolab.2015.08.018>.

- (41) Deng, B.-C.; Yun, Y.-H.; Liang, Y.-Z. Model Population Analysis in Chemometrics. *Chemometrics Intell. Lab. Syst.* **2015**, *149*, 166–176. <https://doi.org/10.1016/j.chemolab.2015.08.018>.
- (42) Jurman, G.; Riccadonna, S.; Furlanello, C. A Comparison of MCC and CEN Error Measures in Multi-Class Prediction. *PLoS One* **2012**, *7* (8), 1–8. <https://doi.org/10.1371/journal.pone.0041882>.
- (43) Boughorbel, S.; Jarray, F.; El-Anbari, M. Optimal Classifier for Imbalanced Data Using Matthews Correlation Coefficient Metric. *PLoS One* **2017**, *12* (6), 1–17. <https://doi.org/10.1371/journal.pone.0177678>.
- (44) Breerton, R. G. Contingency Tables, Confusion Matrices, Classifiers and Quality of Prediction. *J. Chemometrics* **2021**, e3331, 1–5. <https://doi.org/10.1002/cem.3331>.
- (45) Song, J. W.; Lam, S. M.; Fan, X.; Cao, W. J.; Wang, S. Y.; Tian, H.; Chua, G. H.; Zhang, C.; Meng, F. P.; Xu, Z.; Fu, J. L.; Huang, L.; Xia, P.; Yang, T.; Zhang, S.; Li, B.; Jiang, T. J.; Wang, R.; Wang, Z.; Shi, M.; Zhang, J. Y.; Wang, F. S.; Shui, G. Omics-Driven Systems Interrogation of Metabolic Dysregulation in COVID-19 Pathogenesis. *Cell Metab.* **2020**, *32* (2), 188–202.e5. <https://doi.org/10.1016/j.cmet.2020.06.016>.
- (46) Spick, M.; Longman, K.; Frampas, C.; Lewis, H.; Costa, C.; Walters, D. D.; Stewart, A.; Wilde, M.; Greener, D.; Evetts, G.; Trivedi, D.; Barran, P.; Pitt, A.; Bailey, M. Changes to the Sebum Lipidome upon COVID-19 Infection Observed via Rapid Sampling from the Skin. *EClinicalMedicine* **2021**, *33*, 100786. <https://doi.org/10.1016/j.eclinm.2021.100786>.
- (47) Liu, L.; Wei, Q.; Alvarez, X.; Wang, H.; Du, Y.; Zhu, H.; Jiang, H.; Zhou, J.; Lam, P.; Zhang, L.; Lackner, A.; Qin, C.; Chen, Z. Epithelial Cells Lining Salivary Gland Ducts Are Early Target Cells of Severe Acute Respiratory Syndrome Coronavirus Infection in the Upper Respiratory Tracts of Rhesus Macaques. *J. Virology* **2011**, *85* (8), 4025–4030. <https://doi.org/10.1128/jvi.02292-10>.
- (48) Zhang, L.; Xiao, M.; Wang, Y.; Peng, S.; Chen, Y.; Zhang, D.; Zhang, D.; Guo, Y.; Wang, X.; Luo, H.; Zhou, Q.; Xu, Y. Fast Screening and Primary Diagnosis of COVID-19 by ATR-FT-IR. *Anal. Chem.* **2021**, *93* (4), 2191–2199. <https://doi.org/10.1021/acs.analchem.0c04049>.

CAPÍTULO 5

FLORESTA ALEATÓRIA NÃO SUPERVISIONADA COMO ALTERNATIVA PARA RECONHECIMENTO DE PADRÕES E DIAGNÓSTICO DE COVID-19 A PARTIR DE SWAB NASOFARÍNGEO E ESPECTROSCOPIA MIR



- Neste capítulo, uma nova abordagem é proposta para aplicação da floresta aleatória não supervisionada na contribuição dos estudos para diagnóstico de COVID-19.
- Floresta aleatória não supervisionada identificou estruturas nos dados permitindo destacar as variáveis de maior relevância.
- A partir de uma análise não supervisionada foi possível classificar amostras positivas para COVID-19 com sensibilidade de 93,6%.

5. FLORESTA ALEATÓRIA NÃO SUPERVISIONADA COMO ALTERNATIVA PARA RECONHECIMENTO DE PADRÕES E DIAGNÓSTICO DE COVID-19 A PARTIR DE SWAB NASOFARÍNGEO E ESPECTROSCOPIA MIR

Resumo

SARS-CoV-2 é um coronavírus, causador da infecção COVID-19 e de milhões de mortes durante uma pandemia sem precedentes. Uma detecção eficiente e rápida, uso de máscaras e outras medidas de prevenção ainda são fatores cruciais para controle de disseminação do vírus. Espectroscopia na região de infravermelho médio com reflectância total atenuada tem sido proposta como técnica mais rápida e simples para a detecção de infecções virais em biofluidos. Neste trabalho, propomos reconhecimento de padrões pelo uso de floresta aleatória não supervisionada a partir de espectroscopia e *swab* nasofaríngeo. Um total de 1.084 espectros foram adquiridos em quadruplicata de 271 amostras, sendo 156 com diagnóstico Positivo para COVID-19 e 115 com diagnóstico Negativo, ambos confirmados pelo RT-qPCR. A URF foi adaptada para proporcionar um reconhecimento de padrões e extrair informações relevantes na distinção dos grupos. Ademais, é proposto aqui um método de classificação com matriz scores de dimensionamento da URF, pela análise das coordenadas principais, como dado de entrada. A cada conjunto de dados um modelo URF foi desenvolvido com a inserção de *outliers* artificiais, construção de matriz de proximidade e modelo PCoA. A partir do modelo PCoA foi construído um modelo RF com os dados originais e as variáveis preditoras identificadas. Ao final, a matriz de coordenadas principais da URF foi utilizada como entrada de modelo de classificação com os dados separados em treino e teste. Os modelos com o método proposto apresentaram excelentes parâmetros de performance. Um conjunto contendo dois tipos de *swab* nasofaríngeo, com aquisição dos espectros sem procedimento de secagem prévia apresentou modelo URF-PCoA-RF com 93,6% de sensibilidade, 79,4% de especificidade e F1-score de 0,898. Embora o mesmo procedimento de preparo e aquisição de espectros, mas sem variar tipo de *swab*, tenha alcançado maior valor de especificidade (93,3%), considerou-se uma maior aplicabilidade para escolher o

modelo. Essa aplicabilidade potencial é uma importante contribuição para o desenvolvimento de métodos alternativos de triagem no controle de disseminação de COVID-19. Além disso, o método proposto traz uma nova abordagem com maior aproveitamento da URF na extração de informação dos dados.

Palavras-chave: COVID-19; MIR; diagnóstico; Floresta aleatória

5.1. Introdução

No início do ano 2020, a Organização Mundial de Saúde (OMS) declarou como pandêmico o vírus SARS-CoV-2, um tipo de coronavírus, causador da infecção COVID-19. O coronavírus é um vírus de RNA envelopado, fita simples com forma pleomórfica ou esférica com projeções de glicoproteína. Este grupo de vírus foi identificado pela primeira vez na década de 1960 como uma causa de resfriado comum. Entretanto, com surtos denominados de síndrome respiratória severa aguda (SARS) e síndrome respiratória do Oriente Médio (MERS) a comunidade científica observou que esses vírus poderiam causar infecções humanas sérias.^{1,2}

Apesar dos avanços, produção de vacinas, medicamentos e maior conhecimento desde o início desse período pandêmico, ainda se observam altas taxas de internações por conta dessa infecção. Isso ocorre principalmente devido às diferenças no alcance de maior controle e vacinação entre os países, alta disseminação e surgimento de subvariantes. Uma detecção prática e eficiente de sujeitos infectados e demais procedimentos de controle da disseminação ainda são pontos primordiais. Neste contexto, métodos mais rápidos e simples para a detecção de infecções virais em biofluidos têm sido propostos como potencial ferramenta de triagem e combate à propagação dessa doença, que tem causado sérios impactos na saúde e em fatores socioeconômicos.^{1,3}

Atualmente, o método padrão ouro para diagnóstico de COVID-19 é o ensaio de reação em cadeia da transcriptase-polimerase reversa em tempo real (RT-PCR), ou PCR quantitativo (RT-qPCR). Mas este método requer preparação da amostra, o que aumenta o tempo da análise, além do uso de reagentes. Apesar da alta sensibilidade, mesmo o método RT-qPCR, ou os imunoensaios, que também têm sido adotados para diagnóstico, apresentam falsos negativos, o que provoca falha na detecção de pacientes infectados.¹

O uso da espectroscopia vibracional, principalmente na região do infravermelho médio, tem sido proposto como uma técnica com alta potencialidade na busca por um método de diagnóstico a partir de análises de biofluidos. Esta técnica fornece informação de biomoléculas ao nível molecular com as vantagens de menor tempo de análise, não requerer uso de reagentes e preparo mínimo das amostras. Ademais, a dinâmica dos processos inflamatórios, celulares e virais podem provocar alterações bioquímicas que se refletem nas absorções vibracionais das biomoléculas. Isso tem sido abordado associando-se a métodos de estatística multivariada e aprendizagem de máquina para distinção de biofluidos infectados pela COVID-19.⁴⁻¹¹

Em um dos primeiros e mais discutidos artigos com utilização de infravermelho para diagnóstico de COVID-19 em biofluidos, Baraúna *et al.*¹² aplicaram infravermelho na região do médio com acessório de reflectância total atenuada (ATR) para detectar o vírus SARS-CoV-2 a partir de saliva coletada por meio de *swab* orofaríngeo. Os autores distinguiram 70 amostras positivas de 101 negativas por meio de GA-LDA, associando variáveis a processos inflamatórios. O diagnóstico para COVID-19 por meio de MIR também foi proposto por Zhang e colaboradores¹³, mas por meio de análise de outro biofluido, soro sanguíneo. Os autores avaliaram soro sanguíneo de 76 pacientes com infecção causada pelo SARS-CoV-2 e outros vírus respiratórios. Por meio de método de classificação PLS-DA distinguiram 20 amostras de soro saudável, sem sintomas de infecção respiratória, das de soro com infecção por vírus respiratórios comuns, assim como COVID-19. Os autores descreveram algumas mudanças bioquímicas em função da infecção pela COVID-19, expressadas no espectro do biofluido, como mudanças na frequência de absorção de proteínas e lipídios. Entretanto, destacaram o desafio de não alcançar boa distinção de pacientes com COVID-19 assintomáticos e no início da infecção.

Uma caracterização espectroscópica do vírus SARS-CoV-2 foi realizada por Wood *et al.*⁹ utilizando espectroscopia sincrotron-MIR, Raman e outras técnicas. Bandas de absorção de biomoléculas foram descritas como importantes para identificação de constituintes de partículas desse vírus, como proteínas (1657, 1547 e 1517 cm^{-1}) e lipídios (1740, 1464, 1382 e 1341 cm^{-1}). Ademais, os autores aplicaram o método PLS-DA a espectros MIR de 57 amostras de saliva. Outros estudos destacam-se na busca por diagnóstico alternativo de COVID-19 utilizando

diferentes biofluidos, como suspensão em transporte viral¹⁴, saliva^{15,16} e soro¹⁷. Também se destaca na literatura a investigação de mudanças nas frequências de absorções de grupos funcionais de interesse, como lipídios, proteínas e ácidos nucleicos, em soro sanguíneo de indivíduos vacinados para COVID-19.¹⁸ Entretanto, mesmo os estudos mais discutidos apresentam lacunas como pequeno número de amostras, distinção de pacientes assintomáticos, pacientes no início da infecção, outros tipos de infecção, associação de variáveis espectrais, entre outros.

Em trabalho anterior,¹⁶ distinguimos salivas de 138 pacientes infectados pela COVID-19 de 99 não infectados, todos com sintomas de síndromes respiratórias, utilizando espectroscopia MIR associada a diferentes métodos de aprendizagem de máquina. Descrevemos diversas regiões espectrais associadas a alterações bioquímicas em função da presença do vírus SARS-CoV-2 na saliva dos indivíduos. Aqui, neste trabalho, recebemos uma demanda de avaliar a potencialidade de se obter informação discriminante com análise de biofluido nasofaríngeo coletado em *swab*. Dessa forma, este artigo propõe um método de aprendizagem de máquina com potencialidade para extrair informação que contribua para um diagnóstico alternativo para COVID-19 por meio de espectroscopia MIR.

5.2. Metodologia

5.2.1 Amostras

Neste estudo, realizamos análise multivariada associada à espectroscopia MIR, a partir de uma demanda de laboratório clínico na região do Espírito Santo. Recebemos um total de 271 *swab* com material biológico nasofaríngeo coletado por profissionais de saúde das instituições participantes, para avaliar a aplicabilidade de diagnóstico para COVID-19 a partir de um material de uso rotineiro no método padrão, a coleta pelo *swab* nasofaríngeo. Esta coleta foi realizada em pacientes que buscavam os serviços de diagnóstico para COVID-19 pelo método padrão. Após o procedimento de coleta, o material também foi cedido para realização deste estudo, mediante termo de consentimento e esclarecimento dos pacientes e aprovação do comitê de ética da Universidade Federal do Espírito Santo.

Para coleta, realizada por um profissional de saúde, um *swab* de rayon foi inserido em um local equidistante da narina e esfregado suavemente por alguns segundos para absorver as secreções. O *swab* contendo o material biológico foi imediatamente colocado em tubo estéril contendo meio de transporte viral. RT-qPCR foi realizado no Laboratório Central da Secretaria de Saúde do Espírito Santo (LACEN-SESA) para diagnóstico de infecção COVID-19. De um total de 271 pacientes doadores, 156 tiveram diagnóstico Positivo para COVID-19, por meio do método RTq-PCR e 115, diagnóstico Negativo para COVID-19. Dessa forma os materiais coletados, foram identificados e organizados de acordo com seus respectivos diagnósticos e mantidos em freezer (-20 °C) até a aquisição dos espectros. A aprovação ética foi concedida pelo Comitê de Ética da Universidade Federal do Espírito Santo (nº 0993920.1.0000.5071 e nº 31411420.9.0000.8207).

Como preparo de amostras, foi avaliado o procedimento de secagem prévia à aquisição dos espectros. Para isso, cada tubo contendo *swab* com material biológico coletado e fluido de transporte viral (solução salina 0,9% m/v) foi descongelado na véspera da aquisição dos espectros e acondicionado em suporte para secagem à temperatura ambiente, pelo período de uma noite (*overnight*). Para aquisição de espectros, foram utilizados os *swab* diretamente no equipamento, ou seja, sem o procedimento de secagem, e após o procedimento de secagem, configurando-se como conjuntos de dados independentes.

Para análise e diagnóstico COVID-19 tem sido utilizado, de forma padronizada o *swab* de rayon, uma fibra sintética semelhante ao algodão. *Swab* de algodão não deve ser utilizado por ser um interferente na reação realizada pelo método padrão ouro, RT-qPCR. Ao receber os materiais para análise pela espectroscopia MIR, observamos diferenças nos tipos de *swab* utilizados, alguns com a ponta contendo maior quantidade de fibra e haste com espessura maior, sendo denominado aqui como *swab* normal (Figura 5.1, A), enquanto outros apresentavam uma haste mais fina e a ponta com menor quantidade de fibras, denominado aqui como *swab* fino (Figura 5.1, B). Dessa forma, o conjunto de dados teve outra subdivisão em sua organização: *swab* fino (n=190) e *swab* normal (n=81).

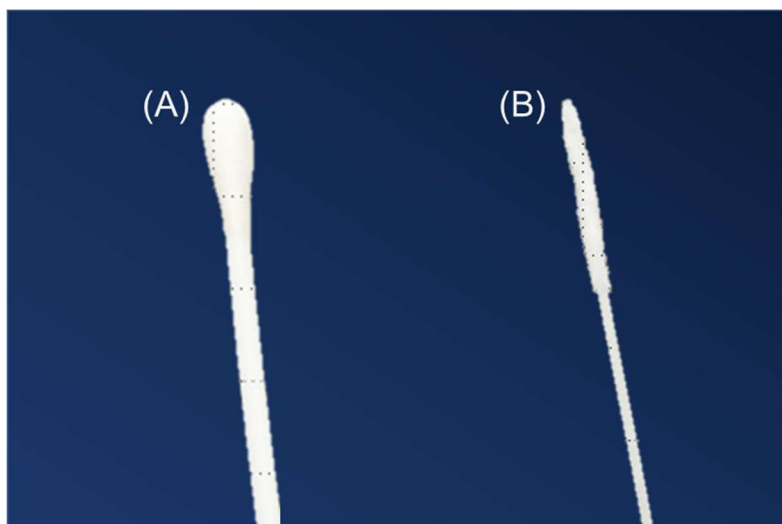


Figura 5.1 Dois tipos de swab de rayon observados neste estudo, sendo utilizados para coleta nasofaríngea para diagnóstico de COVID-19. (A) Swab normal; (B) swab fino.

5.2.2 Espectroscopia MIR

Para aquisição dos espectros, a análise foi realizada com o swab embebido de material biológico posicionado no equipamento, utilizando-se o acessório de amostragem para exercer pressão suficiente e garantir o contato do biofluido presente no swab com o cristal. Um cuidado adicional foi necessário para manter uma padronização de posicionamento do swab no equipamento e não haver contato da haste com o cristal, evitando região com maior risco de contato do cabo com o cristal (Figura 5.2).

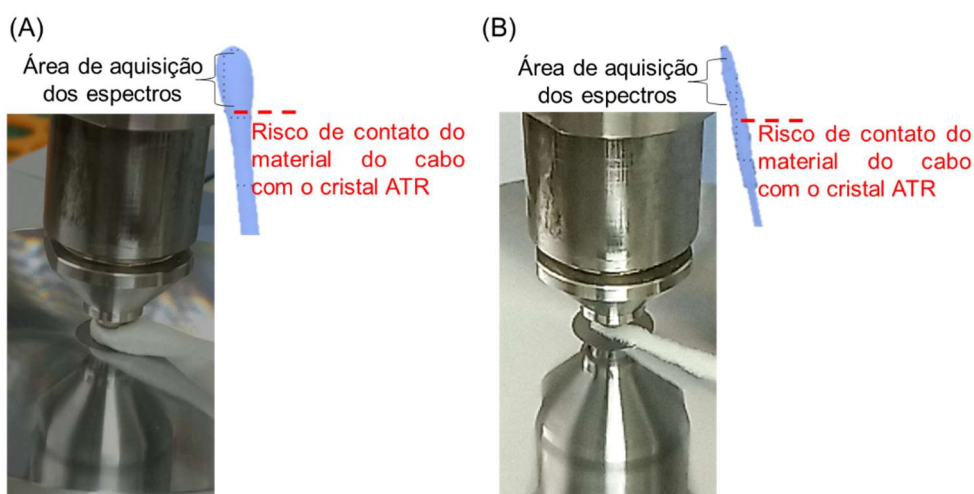


Figura 5.2 Posicionamento no equipamento, dos swab utilizados neste estudo, para evitar contato do material da haste com o cristal ATR. (A) Swab normal; (B) Swab fino.

Os espectros foram adquiridos, utilizando-se o espectrômetro Cary 630 FTIR (Agilent Technologies, Inc.) equipado com um acessório de amostragem por reflectância total atenuada (ATR) com cristal de diamante. A faixa espectral foi de 4.000 a 650 cm^{-1} , no modo absorbância com resolução de 4 cm^{-1} , com 32 *scans* para *background* e amostra. A cada análise, a janela de amostragem de diamante foi limpa com etanol 70% v/v e água Mili Q. O espectro MIR foi medido em quadruplicata, com um tempo médio de 90 segundos para cada amostra, fornecendo um conjunto total de 1.084 espectros. Como houve diferentes tipos de *swab* e diferentes condições de preparo, os espectros adquiridos formaram diversos conjuntos de dados, que após as organizações iniciais, com retiradas de espectros com erros, foram organizados de acordo com a descrição no quadro 5.1. Esses conjuntos de dados foram utilizados de forma independente para as análises multivariadas de aprendizagem de máquina.

Quadro 5.1. Descrição dos conjuntos de dados organizados com os espectros médios (quadruplicata) das amostras de *swab*.

Conjunto de dados	Descrição
01 n=151 (604 espectros) POS=71 (284) / NEG=80 (320)	Conjunto de dados contendo espectros de <i>swab</i> fino com procedimento de secagem.
02 n=190 (760 espectros) POS=90 (360) / NEG=100 (400)	Conjunto de dados contendo espectros de <i>swab</i> fino sem procedimento de secagem.
03 n=72 (288 espectros) POS=61 (244) / NEG=11 (44)	Conjunto de dados contendo espectros de <i>swab</i> normal com procedimento de secagem.
04 n=81 (324 espectros) POS=66 / NEG=15	Conjunto de dados contendo espectros de <i>swab</i> normal sem procedimento de secagem.
05 n=223 (892 espectros) POS=132 (528) / NEG=91 (364)	Conjunto de dados contendo os espectros de todos os <i>swab</i> (normal e fino), com procedimento de secagem.
06 n=271 (1084) POS=156 (624) / NEG=115 (460)	Conjunto de dados contendo os espectros de todos os <i>swab</i> (normal e fino), sem procedimento de secagem.

5.2.3 Análise multivariada dos dados

Todos os espectros foram inicialmente processados para média das quadruplicatas, recorte de extremidades, ajuste de linha de base com algoritmo

airPLS¹⁹ e subtração de espectro médio de swab sem material biológico, adotando-se mesmos procedimentos para armazenagem, secagem e aquisição dos espectros.

Neste estudo, os conjuntos de dados foram submetidos aos modelos de reconhecimento de padrões não supervisionado e supervisionado pelo método floresta aleatória, propondo-se a utilização do método floresta aleatória não supervisionada como ponto de partida para identificação e seleção de variáveis, além de classificação das amostras.

5.2.4 Método floresta aleatória não supervisionada (URF)

O método URF foi utilizado por meio de um algoritmo próprio, adaptado, com geração de *outliers* artificiais aleatórios de mesmo número do conjunto original. RF dos dados originais e sintéticos foi realizada com parâmetro $mtry = \sqrt{\text{variáveis}}$, divisão dos nós pelo índice de impureza de Gini e número total de árvores $nTree = 1000$. Entre esses parâmetros, $nTree$ foi otimizado e escolhido com base na taxa de erro do conjunto OOB. Em dados espectroscópicos, geralmente com alta colinearidade, é necessário $nTree$ suficientemente grande para extrair eficientemente a relação existente entre as variáveis.^{20–23} A análise pelo PcoA foi realizada com base na distância Euclidiana.

5.2.5 URF-RF para identificar variáveis

URF pode ser adotado como um método alternativo para análise exploratória. Neste estudo, aplicamos URF também para identificar e selecionar variáveis relevantes. Para isso, um modelo URF foi construído para identificar se há estrutura nos dados. A partir da projeção das amostras pelos gráficos PCoA é possível visualizar se houve algum tipo de separação entre as amostras cujas classes sejam desconhecidas (amostras não rotuladas), ou sejam previamente conhecidas (amostras rotuladas).

Para identificar as variáveis de maior relevância é possível adotar dois métodos, a depender de conhecimento da classe real das amostras: A - atribuir classe de acordo com a projeção nos gráficos PCoA (amostras não rotuladas); ou, B - atribuir classe de acordo com a classe real (amostras rotuladas). Para a primeira, utiliza-se os valores da PCo que apresentar melhor separação entre os grupos para

atribuição de classes. Um fluxograma das etapas para uso da URF para selecionar variáveis relevantes está apresentada na Figura 5.3.

A partir disso, um modelo RF de classificação binária foi construído somente com os dados originais e suas respectivas classes de interesse. Ao final, utilizou-se o método de identificação de variáveis preditoras e frequência relativa de seleção.²⁰

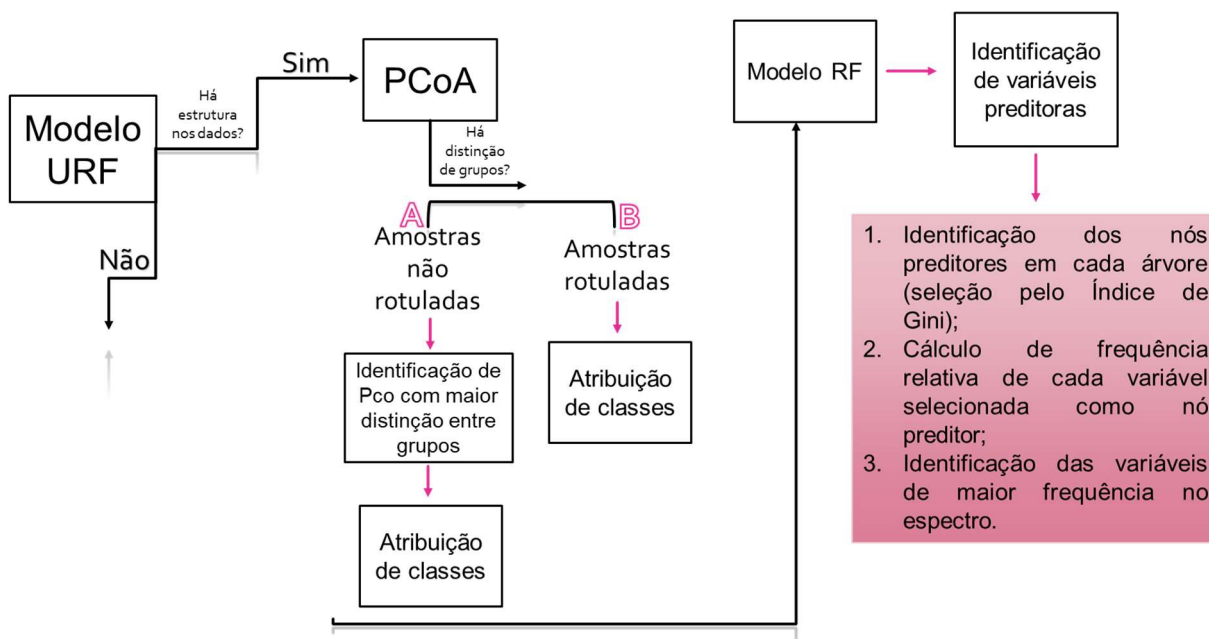


Figura 5.3 Fluxograma de aplicação de URF para selecionar variáveis preditoras.

5.2.6 URF-RF para classificação

URF também está sendo proposto como precursor de método de classificação, neste estudo. No método proposto, um modelo URF é construído. Os dois métodos de atribuição de classes descritos no item 5.2.5 podem ser adotados para atribuição de classes para classificação. Ou seja, amostras rotuladas terão as classes atribuídas de acordo com sua classe real, enquanto amostras não rotuladas terão suas classes atribuídas de acordo com a PCo que apresentar melhor performance na projeção das amostras. Chamaremos neste trabalho de Método A, amostras não rotuladas e Método B para amostras rotuladas.

Com o modelo URF-PCoA construído, utilizamos os valores da matriz de *scores* (coordenadas) do modelo PCoA como entrada para um modelo de classificação (Figura 5.4). Para isso, esses dados foram extraídos do modelo PCoA,

atribuídas as classes e separados em treino (70%) e teste (30%), pelo método Kennard Stone²⁴, garantindo proporção original entre as classes. Um modelo RF de classificação binária foi então construído com o conjunto de treinamento. Ao final da construção das árvores e validação pelos conjuntos OOB, uma validação externa foi realizada com o conjunto de teste externo, que não fez parte da construção do modelo RF. Vale destacar que para se avaliar essa estratégia não foram realizados pré-processamentos nesses dados.

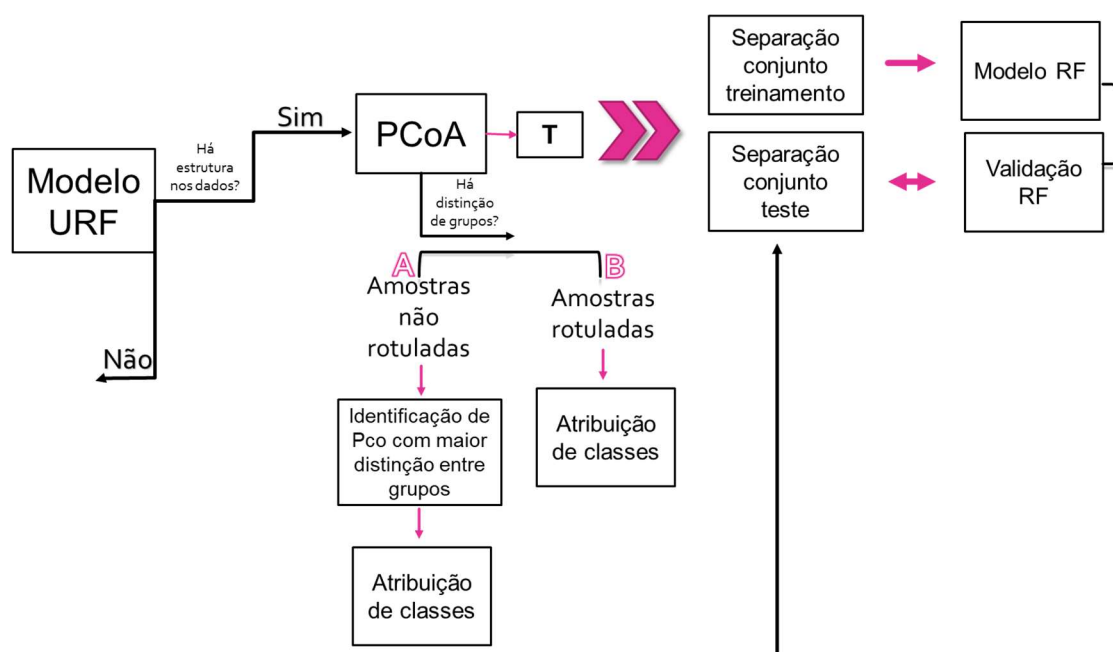


Figura 5.4 Fluxograma de aplicação de URF para classificação de amostras.

Os modelos de classificação foram avaliados pelas métricas: sensibilidade (*Sens*), especificidade (*Espec*), F1-Score e exatidão (*Exat*). Por fim, outras abordagens já consolidadas de classificação foram realizadas nos conjuntos originais de dados para contribuir na avaliação do método proposto, sendo realizado os modelos PLS-DA, GA-LDA e RF nos dados conforme a divisão a seguir: A: variáveis selecionadas pelo URF (item 5.2.5); B: variáveis originais espectro inteiro ($3600 - 900 \text{ cm}^{-1}$); C: variáveis originais região biofingerprint ($1800 \text{ cm}^{-1} - 900 \text{ cm}^{-1}$).

5.3. Resultados e Discussão

Neste estudo, identificamos possíveis estruturas nos dados MIR de *swab* nasofaríngeo para diagnóstico de COVID-19, a partir de modelos não supervisionados de reconhecimento de padrões.

Os perfis espectrais de vários pontos adquiridos de *swabs* vazios estão apresentados na Figura 5.5. Neles há um perfil de absorção de radiação infravermelha pela composição da fibra de rayon, que possui como principal componente a celulose. O perfil espectral deste material apresenta bandas largas e de alta intensidade nas regiões de 1000 a 1200 cm^{-1} (deformação angular de O–H, estiramento de C–O–C e de C–O de carboidratos), que podem suprimir bandas de biomoléculas de interesse. Além dessa região, no perfil espectral de rayon também são esperadas bandas características nas regiões de ~ 3488 - ~ 3447 cm^{-1} (estiramento intramolecular e intermolecular de O–H), ~ 2900 cm^{-1} (estiramento C–H). Ademais, na região considerada importante para obtenção de bandas características de biomoléculas da classe das proteínas (~ 1800 - ~ 1300 cm^{-1}), há algumas bandas características de absorção de grupos funcionais presentes no rayon, como ~ 1423 cm^{-1} (deformação angular de H–C–H e O–C–H), ~ 1335 cm^{-1} (deformação angular no plano de O–H).²⁵ O *swab* do tipo fino (Figura 5.5, A e B) apresentou maior risco de provocar contato da haste com o cristal, consequentemente, observando-se bandas características do polipropileno, material constituinte das hastes de grande maioria dos *swab* comercializados. O *swab* com material mais espesso (Figura 5.5, C e D) tem o perfil espectral mais semelhante ao espectro MIR de *swab* de algodão (Figura 5.5, E e F) e com maior intensidade. Vale ressaltar que para diagnóstico de COVID-19 não é utilizado o *swab* de algodão porque pode interferir na reação realizada no ensaio RT-qPCR. Devido a algumas diferenças entre a fibra natural e a fibra sintética, no perfil espectral do *swab* de algodão há algumas bandas características que não estão presentes no *swab* da fibra rayon, como as bandas em ~ 1735 cm^{-1} (estiramento de C=O de éster), ~ 1460 cm^{-1} (deformação angular no plano de O–H), ~ 1245 cm^{-1} (estiramento de C–C, C–O e C=O) e 1105 cm^{-1} (estiramento assimétrico de C–O–C).²⁵

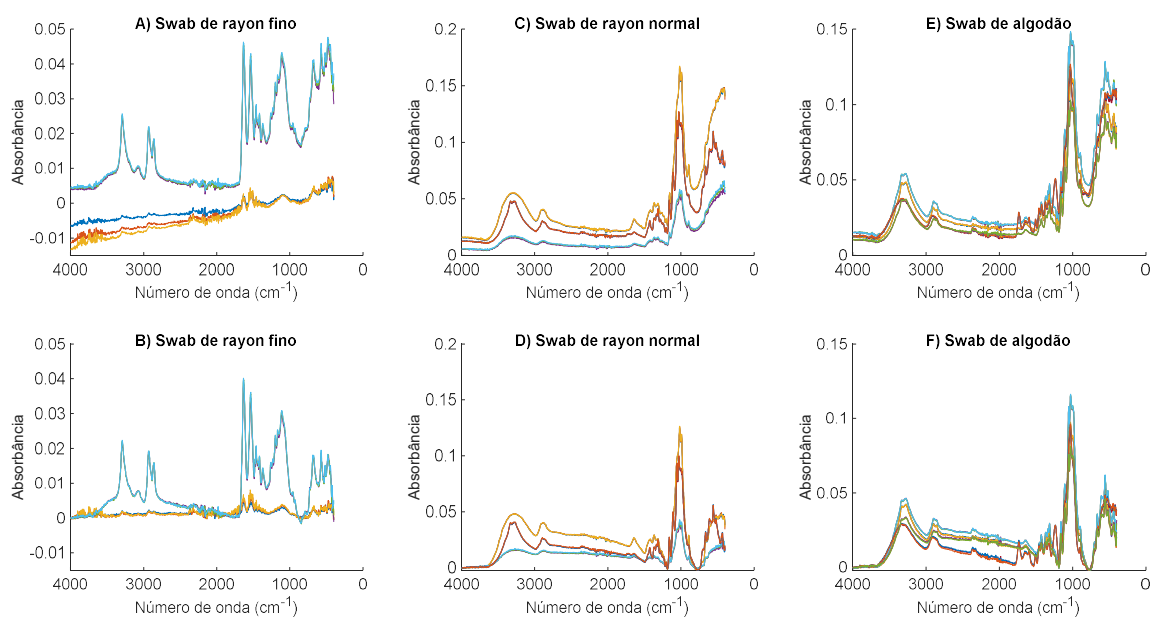


Figura 5.5 Espectros adquiridos em diferentes posições dos swab utilizados neste estudo, sem coleta de material biológico. A e B: Swab de rayon, fino; C e D: Swab de rayon, normal (maior espessura e maior quantidade de fibras; E e F: Swab de algodão. Espectros brutos (A, C e E) *versus* espectros após correção de linha de base (B, D e F).

Essa semelhança nas regiões de bandas características pode ser uma limitação no uso desses dados para diagnóstico de COVID-19 por meio da espectroscopia na região do infravermelho, uma vez que pode haver supressão de bandas importantes para a diferenciação de amostras positivas. A análise desses dados requer uma atenção às fontes de variabilidade e métodos estatísticos mais elaborados para extrair a informação existente, mesmo diante de todas essas interferências. Uma análise exploratória utilizando PCA e URF com os dados sem a subtração de um espectro do *swab* vazio (Figura 5.6) mostrou que a variação do tipo de *swab* utilizado poderia influenciar e se sobrepor à variação de interesse para a classificação. O que se busca extrair é a alteração bioquímica em função da presença do vírus SARS-CoV-2 no biofluido, refletida no espectro, capaz de permitir a classificação e consequente diagnóstico para a infecção COVID-19. Entretanto, pela análise PCA (Figura 5.6, A), a PC1 descreve as amostras pela presença da água com as amostras de *swab* com o líquido do transporte viral, sem o procedimento de secagem, estando posicionadas no sentido de *scores* negativos PC1 <0) sendo descritas principalmente pela região de 3500 - 3000 cm⁻¹, que é conhecida pela absorção de -OH. Já a PC2 descreve o tipo de *swab*, se fino ou se

normal, juntando amostras Positivas e Negativas num mesmo agrupamento. Já pela URF (Figura 5.6, B), há claramente uma distinção em função do tipo de *swab*, separados pela PCo 1. Mas há também uma hierarquia pela PCo2 que distingue, mesmo que com menor valor em distância Euclidiana, os grupos entre Positivos e Negativos. As amostras positivas são projetadas agrupadas no sentido positivo da PCo2, e no sentido negativo as amostras negativas. Esta visualização é importante para compreendermos que há uma estrutura nos dados que promove a distinção entre Positivo e Negativo, mas é menor do que a informação que difere pelo tipo de material do *swab*. Importante destacar que esta análise foi realizada antes da subtração dos espectros do *swab* puro. Este procedimento foi realizado para todos os demais modelos não supervisionados e supervisionados apresentados neste estudo.

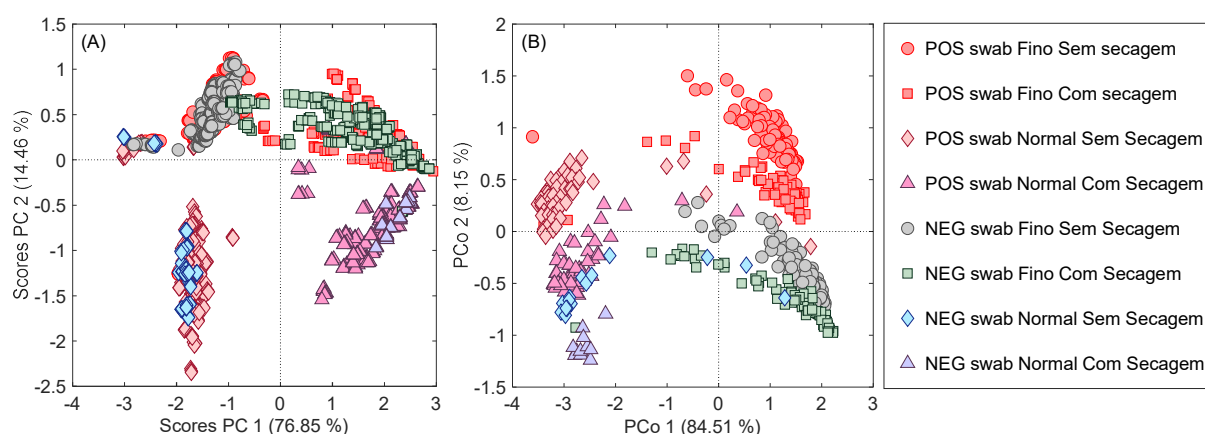


Figura 5.6 Gráficos de análises exploratórias com dados originais sem subtração dos espectros de *swab*. A) Gráfico de Scores PC1 x PC2 de modelo PCA (dados com quadruplicatas, processados com airPLS e centrado na média); B) Gráfico PCoA de URF-PCoA PCo1 x PCo2 (dados com espectros médios, processados com airPLS, MSC e primeira derivada, 7 pontos).

Os espectros dos seis conjuntos de dados organizados e analisados neste estudo são apresentados na Figura 5.7, após subtração do espectro médio do *swab* vazio, bem como processamentos iniciais, como ajuste da linha de base (airPLS)¹⁹ e suavização com filtro Savitzky-Golay²⁶ (SG). É possível observar no perfil dos espectros bandas características e reconhecidas na bioespetroscopia, principalmente na região de biofingerprint, de 1800 a 900 cm^{-1} . Nesta região se observam bandas características de biomoléculas da classe de proteínas, como da ligação amida I ($\sim 1650 \text{ cm}^{-1}$), atribuída a estiramento de C=O e deformação angular C–N; e amida II ($\sim 1550 \text{ cm}^{-1}$), de deformação angular de N–H e estiramento C–N,

além de amida III ($\sim 1.260 \text{ cm}^{-1}$), estiramento C–N).^{4–9} Todas essas regiões são importantes no estudo de amostras infectadas pelo vírus SARS-CoV-2 pela sua característica de envelope com projeções de proteína *spike*, membrana e nucleocapsídeo.^{2,9} Também se observa bandas características de biomoléculas da classe de lipídios, alongamento simétrico C=O ($\sim 1.740 \text{ cm}^{-1}$ - 1720 cm^{-1}), associado a outras bandas ($\sim 1460 \text{ cm}^{-1}$ e 1368 cm^{-1}), carboidratos (alongamento simétrico CO–O–C a $\sim 1.155 \text{ cm}^{-1}$), ácido nucleico (estiramento assimétrico de grupo fosfato, $\sim 1.235 \text{ cm}^{-1}$, e simétrico $\sim 1.080 \text{ cm}^{-1}$). Entretanto, espectros obtidos com *swab* mais espesso (Figura 5.7, E, F, G e H) apresentaram supressão na região de banda característica de lipídios ($\sim 1.740 \text{ cm}^{-1}$ - 1720 cm^{-1}). Além disso, espectros de amostras negativas em *swab* normal com procedimento de secagem (Figura 5.7, F) apresentam-se com baixa intensidade em comparação aos demais espectros. Os espectros desse conjunto de dados antes e após a subtração dos espectros de swab estão apresentados no material suplementar (Capítulo 7, item 7.3, Figura S01).

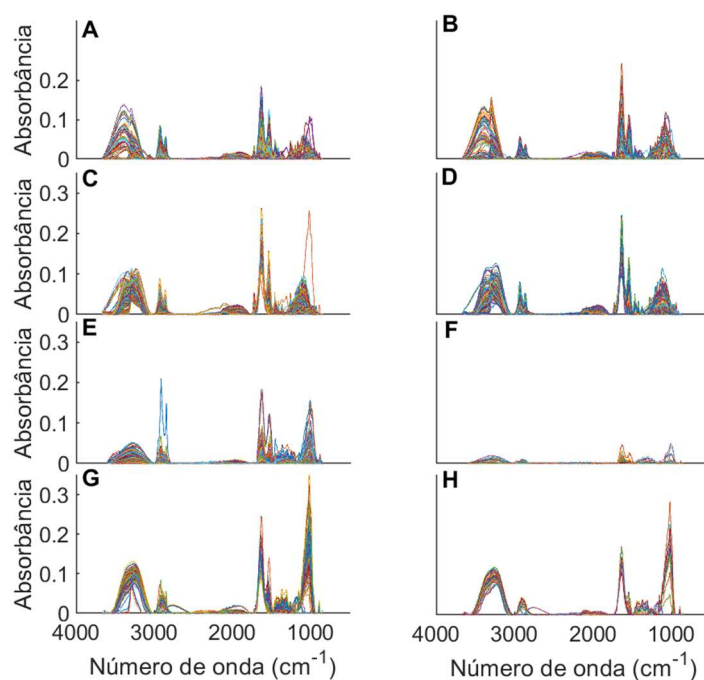


Figura 5.7 Espectros dos conjuntos de dados após pré-processamentos iniciais (ajuste de linha de base, subtração de espectro, suavização SG). A) Conjunto de amostras Positivo COVID-19, Swab fino com secagem (n=71); B) Conjunto de amostras Negativo COVID-19, Swab fino com secagem (n=80); C) Conjunto de amostras Positivo COVID-19, Swab fino sem secagem (n=90); D) Conjunto de amostras Negativo COVID-19, Swab fino sem secagem (n=100); E) Conjunto de amostras Positivo COVID-19, Swab normal com secagem (n=61); F) Conjunto de amostras Negativo COVID-19, Swab normal com secagem (n=11); G) Conjunto de amostras Positivo COVID-19, Swab normal sem secagem (n=66); H) Conjunto de amostras Negativo COVID-19, Swab normal sem secagem (n=60).

Além disso, mesmo os espectros dos *swab* após procedimento de secagem (Figura 5.7 A, B, E e F) apresentam em alguns deles bandas características (largas) na região de 3400 cm^{-1} , ($-\text{OH}$), indicando que o procedimento de secagem não foi homogêneo, gerando uma fonte de variação indesejada nos dados. Os espectros médios normalizados estão apresentados na Figura 5.8, onde não é possível identificar diferenças discrepantes nos perfis espectrais de amostras positivas *versus* negativas, mas algumas diferenças nas intensidades. O procedimento de secagem contribui para expressar as bandas características dos grupos funcionais Amida I ($\sim 1650\text{ cm}^{-1}$) e Amida II ($\sim 1550\text{ cm}^{-1}$), atribuídas a proteínas. Pelo aspecto do vírus causador da infecção COVID-19, é esperado que as diferenças de níveis de proteínas entre as amostras infectadas e não infectadas reflitam-se nesta, podendo ser associada a um aumento de proteínas virais ou anticorpos produzidos pela pessoa infectada.²⁷

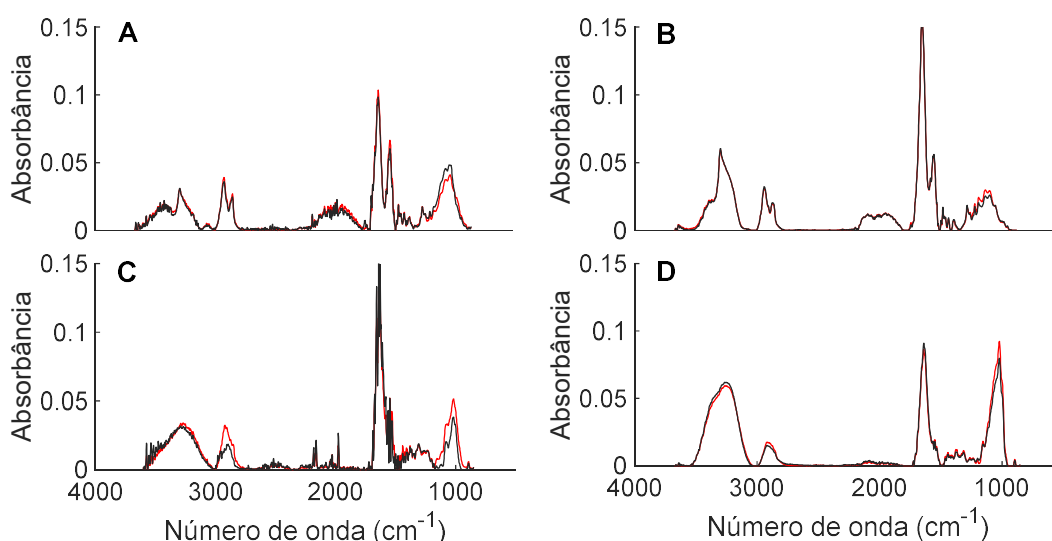


Figura 5.8 Espectros médios Positivo (em vermelho) e Negativo (em cinza) dos conjuntos de dados após pré-processamentos iniciais (ajuste de linha de base, subtração de espectro, suavização SG). A) Conjunto *swab* fino com secagem; B) Conjunto *swab* fino sem secagem; C) Conjunto *swab* normal com secagem; D) Conjunto *swab* normal sem secagem.

O método URF foi utilizado como uma análise alternativa de reconhecimento de padrões não supervisionado. Por meio deste método foi possível identificar estruturas existentes em cada conjunto de dados, assim como extrair as variáveis de maior relevância na distinção dos grupos. Excetuando-se do conjunto de *swab* normal com secagem (Figura 5.9, C), os demais conjuntos apresentam uma

estrutura. Entretanto, a estrutura apresentada não está indicando uma potencialidade discriminante, o que seria esperado a partir de uma maior distância na coordenada principal 1 (PCo 1). Ao invés disso, foi possível observar tendências de discriminação nas PCo 1 x PCo 4 (total de 93,61% de variância de PCo 1 a PCo4), no conjunto de *swab* fino com secagem (Figura 5.9 A) e PCo 2 x PCo 4 (39,89% de variância até PCo4) no de *swab* normal sem secagem ((Figura 5.9 D). Enquanto no conjunto de *swab* fino sem procedimento de secagem ((Figura 5.9 B) essa tendência foi observada pelas PCo 2 x PCo 3 (48,99% de variância). Ademais, é possível observar que o modelo nos conjuntos com procedimento de secagem descreve maior variância dos dados.

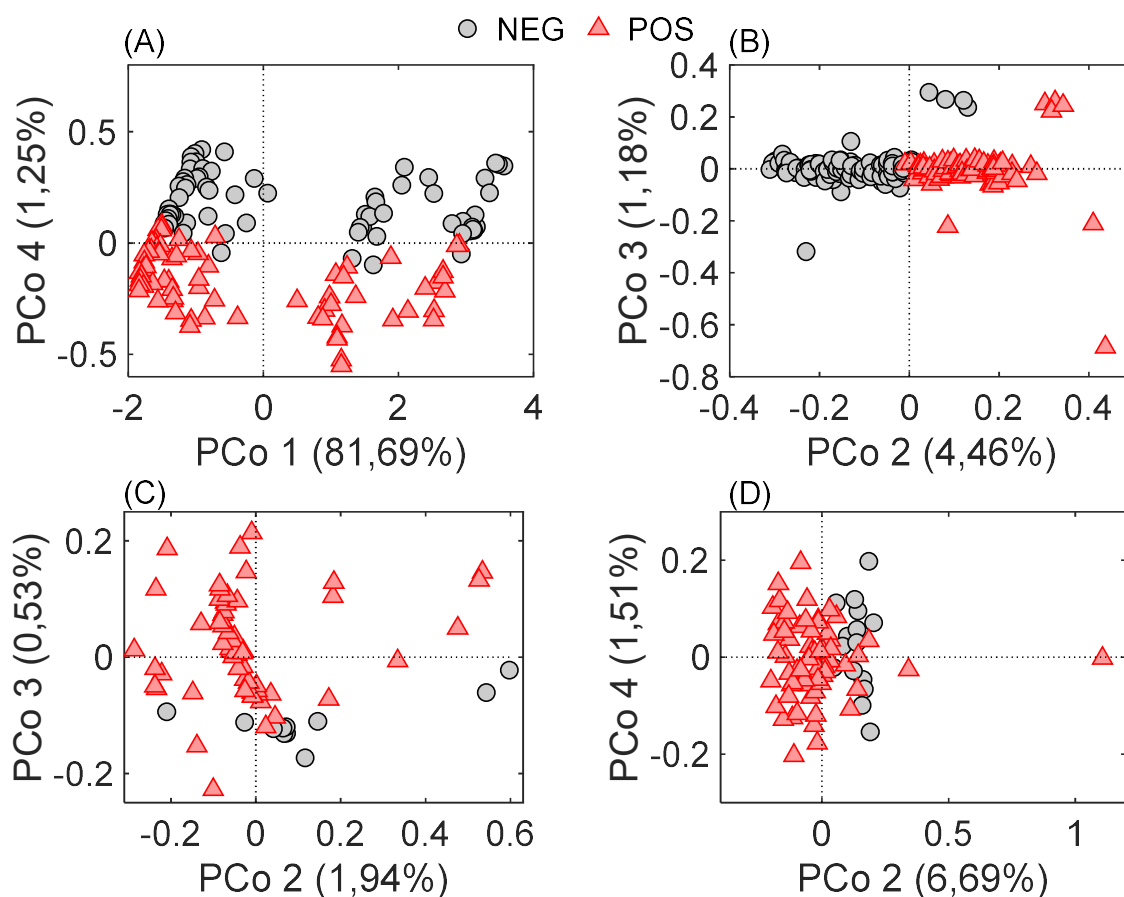


Figura 5.9 Gráficos de coordenadas principais, a partir de modelo URF-PCoA (distância Euclidiana) de conjuntos de dados avaliados. A) Conjunto *swab* fino com secagem; B) Conjunto *swab* fino sem secagem; C) Conjunto *swab* normal com secagem; D) Conjunto *swab* normal sem secagem.

Com o objetivo de analisar a potencialidade discriminante dos dados independentemente do tipo de *swab* a ser utilizado, também realizamos a análise

pelo URF a partir de uma junção desses conjuntos de dados. Esta análise foi realizada considerando-se um contexto de aplicação do método em grande demanda, no qual não seria possível o controle efetivo do tipo de *swab* a ser realizado para a coleta. Dessa forma, a junção dos conjuntos *swab* fino e *swab* normal resultou em dois novos modelos URF, cujos conjuntos totais são: com procedimento de secagem (Figura 5.10 A) e sem procedimento de secagem (Figura 5.10 B), ambos pré-processados pelos métodos SNV seguido de 1ª derivada (janela 9 pontos, polinômio 2º grau). Como esperado, os resultados de cada conjunto são refletidos nestes modelos a partir da junção dos dados. Observa-se uma maior variância do conjunto cujo procedimento de secagem foi adotado (Figura 5.10 A), com um total de 96,04% das 5 primeiras PCo, em comparação a 61,10%, também das 5 PCo.

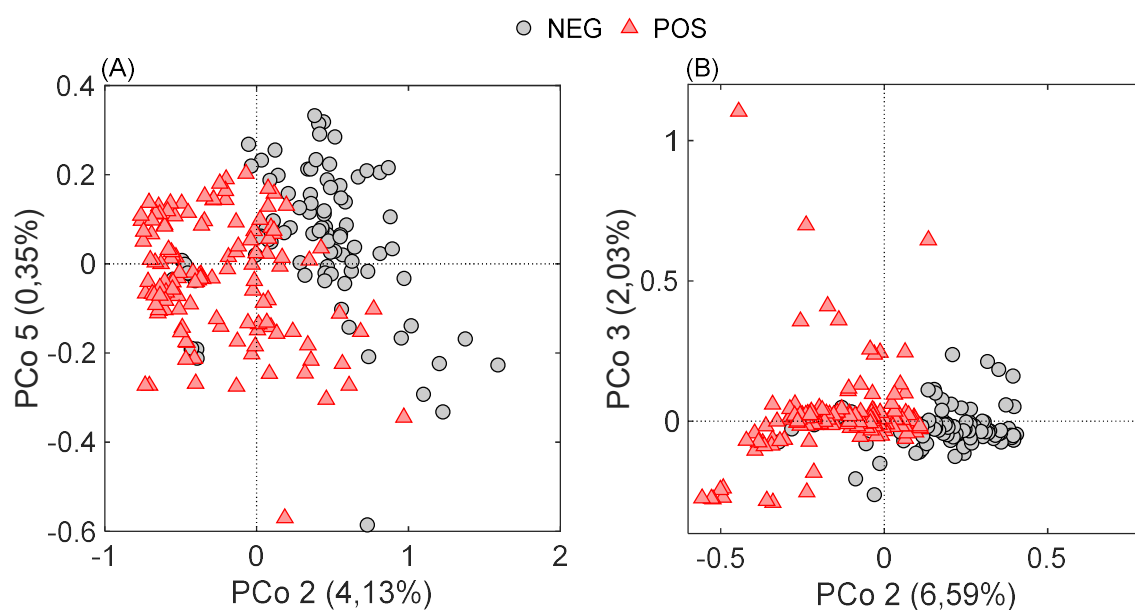


Figura 5.10 Gráficos de coordenadas principais, a partir de modelo URF-PCoA (distância Euclidiana) de conjuntos de dados avaliados. A) Conjunto *swab* fino + *swab* normal com secagem; B) Conjunto *swab* fino + *swab* normal sem secagem.

5.3.1 Seleção de Variáveis a partir de URF

Neste estudo, foram propostos dois métodos para identificação de variáveis de maior relevância a partir de uma análise pela URF. Isso se fez importante neste contexto de uma análise exploratória na busca pela extração de informação que contribua para distinção de amostras infectadas pelo vírus SARS-CoV-2. Apresentamos aqui uma comparação dessa identificação pelos dois métodos a partir

da URF-PCoA dos conjuntos de dados *swab* fino com secagem (01) e *swab* fino sem secagem (02), utilizando a região de *fingerprint* (900 – 1800 cm^{-1}). Estes conjuntos foram escolhidos, pois apresentam condições de realização de ambos os métodos propostos.

Vale destacar que o método de identificação das variáveis com a utilização de amostras sem rótulo de classe, utilizando-se o resultado da análise PCoA e as distinções observadas no gráfico de coordenadas principais, parte do pressuposto de uma análise crítica frente ao resultado gráfico.

É possível observar que para este conjunto de dados as regiões selecionadas são equivalentes, com algumas alterações na concentração de frequência, em que o método pela atribuição de classes a partir da projeção das amostras pelo gráfico de PCo (Figura 5.11 A), apresenta menores pontos com concentrações, em comparação ao método 02, que tem as classes atribuídas a partir do rótulo de classe real das amostras (Figura 5.11 B). Isso pode ser explicado pelo fato de que ao se utilizar uma rotulagem estimada por meio do gráfico de PCoA o modelo pode relacionar erroneamente informação espectral com alguma variação nos dados, ou, ao que parece ter ocorrido, diluir-se variações entre os grupos, ao se incluir variação intra-grupo.

As regiões espectrais de maior relevância ($> \frac{1}{4}$ frequência relativa) de acordo com a análise pela URF-PCoA apresentados na Figura 5.11 (A e B) são as regiões de $\sim 1735 \text{ cm}^{-1}$, $\sim 1697 \text{ cm}^{-1}$, $\sim 1573 - \sim 1588 \text{ cm}^{-1}$, $\sim 1390 - 1400 \text{ cm}^{-1}$; $\sim 1280 - 1360 \text{ cm}^{-1}$; $\sim 985 - 1000 \text{ cm}^{-1}$ e $\sim 937 \text{ cm}^{-1}$. Vale destacar que algumas dessas regiões espectrais ($\sim 1573 - \sim 1588 \text{ cm}^{-1}$ e $\sim 985 - 1000 \text{ cm}^{-1}$) apresentaram frequência relativa significativa apenas no modelo pelo método de amostras rotuladas.

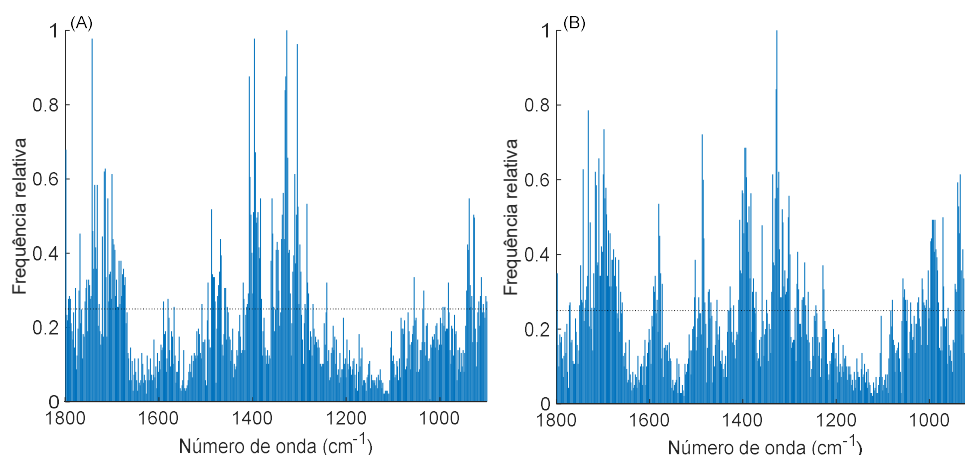


Figura 5.11 Demonstração gráfica da frequência relativa para identificação de variáveis preditoras, a partir de dois métodos de identificação: A) modelo URF-PCoA amostras não rotuladas (conjunto *swab* fino com secagem); B) modelo URF-PCoA amostras rotuladas (conjunto *swab* fino com secagem).

As variáveis de maior frequência relativa no conjunto de amostra a partir do *swab* fino sem procedimento de secagem (conjunto 02) são: $\sim 1748 - 1711 \text{ cm}^{-1}$; $\sim 1655 \text{ cm}^{-1}$; $\sim 1595 - 1560 \text{ cm}^{-1}$; $\sim 1490 \text{ cm}^{-1}$; $1405 - 1383 \text{ cm}^{-1}$; $\sim 1350 - 1269 \text{ cm}^{-1}$; $\sim 1083 \text{ cm}^{-1}$; $\sim 1029 - 989 \text{ cm}^{-1}$; $940 - 920 \text{ cm}^{-1}$ (Figura 5.12). Nesta seleção observa-se uma menor concentração de variáveis com frequência significativa e uma maior distribuição de frequência por toda a região espectral.

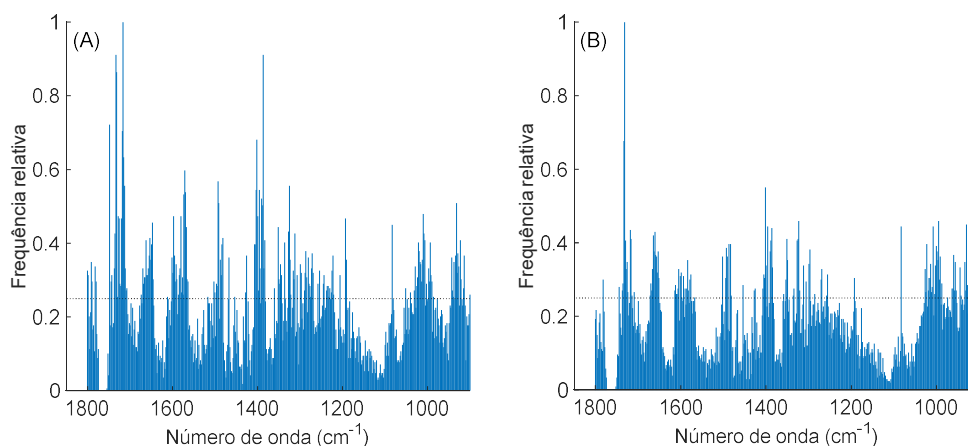


Figura 5.12 Demonstração gráfica da frequência relativa para identificação de variáveis preditoras, a partir de dois métodos de identificação: A) modelo URF-PCoA amostras não rotuladas (conjunto *swab* fino sem secagem); B) modelo URF-PCoA amostras rotuladas (conjunto *swab* fino sem secagem).

Em todos os conjuntos analisados pelo método proposto foi possível destacar as informações de biomoléculas importantes para o estudo das alterações bioquímicas relacionadas ao vírus, como bandas características de lipídios (1745 cm^{-1}

¹, 1660 cm⁻¹, 1490 cm⁻¹ e 1390 cm⁻¹), regiões características dos grupos funcionais Amida I (1650 cm⁻¹), Amida II (1500 cm⁻¹) e Amida III (1400 – 1300 cm⁻¹), além de banda característica de grupos fosfodiéster de ácidos nucleicos (1083 cm⁻¹). Essas informações estão de acordo com bandas destacadas por Wood e colaboradores na caracterização de amostras de saliva infectadas e não infectadas pela COVID-19.²⁷

5.3.2 Classificação das amostras a partir de URF

Neste estudo, a classificação pelo método URF-RF foi realizada como uma proposta de melhor aproveitamento da URF como método alternativo de reconhecimento de padrões. São apresentados na Tabela 5.1 os parâmetros de desempenho dos modelos RF construídos a partir de modelo URF-PCoA, com a atribuição de classes pelos métodos A (amostras não rotuladas) e B (amostras rotuladas), descritos no item 5.4.2.

Tabela 5.1 Parâmetros de desempenho de modelos de classificação URF-PCoA-RF, para a classe Positivo, de conjuntos de treinamento e teste (linha sombreada).

Conjunto dados	Modelo	Método	Pré-processamento	PCo	SENS (%)	ESPEC (%)	F-Score	EXAT (%)
01	URF/Rf	A	----	4	82,8	95,8	0,889	88,7
					90,9	91,3	0,909	91,1
		B	----	----	92,0	83,9	0,876	87,7
02	URF/Rf	A	----	2	85,7	79,2	0,818	82,2
					96,8	87,1	0,917	91,7
		B	----	----	92,6	90,0	0,909	91,2
03	URF/Rf	A	----	4	82,5	95,7	0,881	89,5
					77,8	93,3	0,842	86,0
		B	----	----	58,5	88,9	0,727	64,0
04	URF/Rf	A	----	1	60,0	100	0,750	63,6
					---	---	---	---
		B	----	----	83,7	92,9	0,900	96,9
05	URF/Rf	A	----	2	60,9	100	0,757	62,5
					---	---	---	---
		B	----	----	80,9	67,2	0,787	75,0
06	URF/Rf	A	----	2	81,4	70,8	0,823	77,6
					85,9	51,6	0,782	71,8
		B	SNV / 1ª DERIVADA (9)	----	90,0	59,3	0,828	77,6
06	URF/Rf	A	----	2	67,5	77,3	0,740	71,4
					80,9	82,5	0,819	81,7
		B	----	----	96,3	80,2	0,913	89,5
					93,6	79,4	0,898	87,6

A = Método com estimativa de atribuição de classes pelo gráfico PCoA (amostras não rotuladas). B

= Método com atribuição de classe real (amostras rotuladas).

SENS = Sensibilidade; ESPEC = Especificidade; EXAT = Exatidão.

Para o conjunto de dados 01 (*swab* fino com procedimento de secagem), o método proposto com a classificação por meio de URF-PCoA-RF apresentou ganho, uma vez que os modelos desenvolvidos com as variáveis originais apresentaram sensibilidade de 50 a 87,5 % e especificidade de 43,7 a 76,2% (Tabela S01), para o conjunto de teste externo, inferiores aos valores de 85,7% e 90,9%, sensibilidade e 79,2 e 91,3, de especificidade, obtidos com a classificação URF-PCoA-RF.

Para o conjunto de dados 02 (*swab* fino sem procedimento de secagem), o método proposto também apresentou ganho de informação. Com os modelos PLSDA, GALDA e RF, foram obtidos os valores de sensibilidade de 33,3 a 90% e especificidade de 51,8 a 85,2%. Entretanto, apesar dos parâmetros de performance próximos, o conjunto de dados de *swab* fino sem procedimento de secagem proporcionou melhor resultado nos modelos URF/RF (F1-score de 0,909 e 0,842, métodos A e B, respectivamente), além de apresentar melhor aplicabilidade.

Não foi possível obter modelos URF-PCoA-RF pelo método B para os conjuntos 03 e 04 (*swab* normal). Em contrapartida, foi possível um bom modelo GALDA com esses mesmos dados, porém com redução do conjunto de treinamento para diminuir a discrepância entre as classes. Ademais, a junção de todos os conjuntos com diferentes tipos de *swab*, reduziu os parâmetros, principalmente de especificidade (Tabela 5.1 e Tabela S02). Essa redução é mais visível no conjunto 05, dados com procedimento de secagem.

Foram construídos modelos pelo método proposto utilizando somente a região de *biofingerprint* para construção dos modelos URF-PCoA. Entretanto, não houve melhora nos resultados (Tabela S03, Figuras S02 e S03). Entre os demais modelos realizados (Tabela S01), o método GA-LDA a partir da região *fingerprint* obteve melhores performances. Vale destacar que para os diversos modelos apresentados na Tabela S01, os modelos PLSDA apresentaram maiores números de amostras identificadas como anômalas pelos testes $Q_{residual} \times T^2_{Hotelling}$, principalmente nos conjuntos de dados 01 e 05 (com procedimento de secagem). O número de amostras em cada subconjunto, bem como número de *outliers* identificados por modelo, estão apresentados no material suplementar (Tabelas S04 e S05). Vale ressaltar que para ambos os modelos de classificação RF quanto os modelos URF-PCoA-RF, os parâmetros *nTrees*, *mtry* e número mínimo de nó folha foram otimizados (Figuras S04e S05).

Os resultados da classificação pelo método proposto neste trabalho são equivalentes ou superiores a algumas abordagens semelhantes. O uso de análises com medição por ATR-IR diretamente de *swab* contendo biofluido foi relatado inicialmente por Barauna e colaboradores.¹² Eles aplicaram GA-LDA a 171 espectros de biofluido orofaríngeo coletado por *swab* de algodão, obtendo sensibilidade de 95% e especificidade de 89% na validação do modelo.

O uso de uma suspensão contendo biofluido de coleta de *swab* nasofaríngeo foi avaliada por Nogueira e colaboradores²⁸ para diagnóstico por meio de ATR-IR. Em estudo contendo 65 pacientes com suspeita de COVID-19, cuja coleta foi realizada com o uso de um tipo de meio de transporte viral, e 178 com coleta realizada com outro tipo de meio de transporte viral, os autores obtiveram modelo de classificação com sensibilidade de 84% e 87% e especificidade de 66% e 64%, para cada transporte viral avaliado, respectivamente. Recentemente, Heino e colaboradores²⁹ também apresentaram uma avaliação dessa metodologia em um coorte maior (558 amostras positivas e 558 negativas). Os autores apresentaram resultados médios de validação cruzada (k-fold) com diferentes configurações de modelo PLS-DA a partir dos espectros recortados na região de *fingerprint*, com valores médios 61, 64 e 63%, de sensibilidade, especificidade e exatidão, respectivamente.

Neste trabalho apresentamos um método de aprendizagem de máquina que proporciona ganho de informação nesta análise, adquirindo modelos com sensibilidade em torno de 90%. Entretanto, foram demonstrados aqui algumas limitações ao se empregar a análise espectroscópica utilizando o *swab*. A principal limitação é a possibilidade de incluir fonte de variação indesejada. Essa variação indesejada pode ser originada de diversos fatores, desde a interferência do material da fibra do *swab*, como observado, até uma baixa concentração de partículas virais na amostra. Este último pode estar associado a fatores externos associados à coleta. Wyle e colaboradores³⁰ relataram maior quantidade de cópias de RNA na saliva do que nos fluidos nasofaríngeos. Dessa forma, considerando-se que a combinação de modelos de classificação a partir de saliva, apresentado no Capítulo 4, obteve um bom desempenho na classificação de amostras positivas (93% de sensibilidade, 74,0% de especificidade e 83% de precisão), a modelagem com *swab* apresentou performance equivalente, porém com maior risco de variabilidade indesejada além

da necessidade de métodos quimiométricos mais aprimorados para uma extração eficiente da informação contida nos dados. Ademais, apesar de ter sido possível a elaboração de modelo de classificação com dados completos incluindo diferentes tipos de *swab*, houve maior índice de *outliers* para modelos PLS-DA e GA-LDA e perda de performance de modelos URF/RF. Além disso, apesar de este estudo ser realizado para atender a uma demanda das instituições de saúde, não foi possível receber dados sociodemográficos e clínicos dos pacientes doadores para avaliação de possíveis influências de outros fatores.

O método RF já foi aplicado neste tipo de problema. Guleken e colaboradores³¹ obtiveram excelentes resultados (Sens > 90%) com método supervisionado RF para identificar e associar espectros ATR-IR de soro sanguíneo a mudanças em parâmetros bioquímicos de mulheres grávidas infectadas por COVID-19. O uso de URF também identificou estrutura em dados de ATR-IR de saliva em estudo recente sobre COVID-19.¹⁶

Afanador e colaboradores²³ descreveram a potencialidade do método URF nos estudos na área biológica, destacando ainda que este método não seria tão aproveitado nesta área pela dificuldade de se obter informação das variáveis importantes. Neste trabalho, apresentamos a possibilidade de utilizar a URF como análise exploratória e como precursora para extrair informações relevantes. Apresentamos ainda a associação da URF a modelos de classificação binária pela RF para classificação das amostras, mesmo que se esteja diante de uma investigação inicial, ou seja, sem o conhecimento das classes.

5.4. Conclusões

Floresta aleatória não supervisionada foi proposta neste trabalho como um método alternativo de reconhecimento de padrões, incluindo-se análise não supervisionada, seleção de variáveis e classificação para diagnóstico de COVID-19 a partir de biofluido nasofaríngeo coletado em *swab*. Pelo método proposto é possível identificar que há estrutura nos dados analisados. Essa estrutura permite distinguir amostras positivas de negativas com alta exatidão. Entretanto, a utilização dessa metodologia empregando o *swab* necessita de maior investigação das fontes de variabilidade indesejadas. O método proposto permite uma classificação utilizando diferentes tipos de *swab* com exatidão de 87,6%, embora essa não-

padronização seja mais suscetível a erros aleatórios e isso possa provocar maior fonte de variação indesejada. Portanto, a proposta de utilização de biofluidos coletados de diferentes *swab*, aquisição de espectro direto no equipamento, sem adição de tempo de secagem, visa a uma alta aplicabilidade. Isso é uma importante contribuição no contexto de triagem de indivíduos infectados.

Agradecimentos

Os autores agradecem aos pacientes doadores e a todos os profissionais das instituições de saúde participantes; NCQP/LABPETRO-UFES, FAPES, CAPES e CNPq.

Financiamento: Este estudo foi apoiado pela CAPES (Código Financeiro 001); FAPES (processo número 33530.503.20537.12092017; nº 151/2020, nº 165/2021 e nº 691/2022) e CNPq (processo nº 422515/2016-7; #401870/2020-0 e #310057/2020-5).

5.5. Referências

- (1) Khan, R. S.; Rehman, I. U. Spectroscopy as a Tool for Detection and Monitoring of Coronavirus (COVID-19). *Expert Review of Molecular Diagnostics*. Taylor and Francis Ltd July 2, 2020, pp 647–649. <https://doi.org/10.1080/14737159.2020.1766968>.
- (2) Ke, Z.; Oton, J.; Qu, K.; Cortese, M.; Zila, V.; McKeane, L.; Nakane, T.; Zivanov, J.; Neufeldt, C. J.; Cerikan, B.; Lu, J. M.; Peukes, J.; Xiong, X.; Kräusslich, H. G.; Scheres, S. H. W.; Bartenschlager, R.; Briggs, J. A. G. Structures and Distributions of SARS-CoV-2 Spike Proteins on Intact Virions. *Nature* **2020**, 588 (7838), 498–502. <https://doi.org/10.1038/s41586-020-2665-2>.
- (3) Mir, J. M.; Khan, M. W.; Shalla, A. H.; Maurya, R. C. A Nonclinical Spectroscopic Approach for Diagnosing Covid-19: A Concise Perspective. *J Appl Spectrosc* **2021**, 88 (4), 765–771. <https://doi.org/10.1007/s10812-021-01238-9>.
- (4) Santos, M. C. D.; Morais, C. L. M.; Nascimento, Y. M. Spectroscopy with Computational Analysis in Virological Studies: A Decade (2006e2016). *Trends in Analytical Chemistry* **2017**, 97 (January), 244–256. <https://doi.org/10.1016/j.trac.2017.09.015>.

- (5) Movasaghi, Z.; Rehman, S.; Rehman, I. U. Fourier Transform Infrared (FTIR) Spectroscopy of Biological Tissues. *Appl. Spectrosc. Rev.* **2008**, *43* (2), 134–179. <https://doi.org/10.1080/05704920701829043>.
- (6) Morais, C. L. M.; Lima, K. M. G.; Singh, M.; Martin, F. L. Tutorial: Multivariate Classification for Vibrational Spectroscopy in Biological Samples. *Nat. Protoc.* **2020**, *15* (7), 2143–2162. <https://doi.org/10.1038/s41596-020-0322-8>.
- (7) Butler, H. J.; Smith, B. R.; Fritzsche, R.; Radhakrishnan, P.; Palmer, D. S.; Baker, M. J. Optimised Spectral Pre-Processing for Discrimination of Biofluids via ATR-FTIR Spectroscopy. *Analyst* **2018**, *143* (24), 6121–6134. <https://doi.org/10.1039/c8an01384e>.
- (8) Mitchell, A. L.; Gajjar, K. B.; Theophilou, G.; Martin, F. L.; Martin-Hirsch, P. L. Vibrational Spectroscopy of Biofluids for Disease Screening or Diagnosis: Translation from the Laboratory to a Clinical Setting. *J. Biophotonics* **2014**, *7* (3–4), 153–165. <https://doi.org/10.1002/jbio.201400018>.
- (9) Wood, B. R.; Kochan, K.; Bedolla, D. E.; Salazar-Quiroz, N.; Grimley, S. L.; Perez-Guaita, D.; Baker, M. J.; Vongsivut, J.; Tobin, M. J.; Bambery, K. R.; Christensen, D.; Pasricha, S.; Eden, A. K.; Mclean, A.; Roy, S.; Roberts, J. A.; Druce, J.; Williamson, D. A.; McAuley, J.; Catton, M.; Purcell, D. F. J.; Godfrey, D. I.; Heruad, P. Infrared Based Saliva Screening Test for COVID-19. *Angew. Chem. Int. Ed. Engl.* **2021**, *60* (31), 17102–17107. <https://doi.org/10.1002/anie.202104453>.
- (10) Mir, J. M.; Khan, M. W.; Shalla, A. H.; Maurya, R. C. A Nonclinical Spectroscopic Approach for Diagnosing Covid-19: A Concise Perspective. *J Appl Spectrosc* **2021**, *88* (4), 765–771. <https://doi.org/10.1007/s10812-021-01238-9>.
- (11) Santos, M. C. D.; Morais, C. L. M.; Lima, K. M. G. ATR-FTIR Spectroscopy for Virus Identification: A Powerful Alternative. *Biomed Spectrosc Imaging* **2020**, *9* (3–4), 103–118. <https://doi.org/10.3233/bsi-200203>.
- (12) Barauna, V. G.; Singh, M. N.; Barbosa, L. L.; Marcarini, W. D.; Vassallo, P. F.; Mill, J. G.; Ribeiro-Rodrigues, R.; Campos, L. C. G.; Warnke, P. H.; Martin, F. L. Ultrarapid On-Site Detection of SARS-CoV-2 Infection Using Simple ATR-FTIR Spectroscopy and an Analysis Algorithm: High Sensitivity and Specificity. *Anal. Chem.* **2021**, *93* (5), 2950–2958. <https://doi.org/10.1021/acs.analchem.0c04608>.
- (13) Zhang, L.; Xiao, M.; Wang, Y.; Peng, S.; Chen, Y.; Zhang, D.; Zhang, D.; Guo, Y.; Wang, X.; Luo, H.; Zhou, Q.; Xu, Y. Fast Screening and Primary Diagnosis of COVID-19 by ATR-FT-IR. *Anal Chem* **2021**, *93* (4), 2191–2199. <https://doi.org/10.1021/acs.analchem.0c04049>.
- (14) Nogueira, M. S.; Leal, L. B.; Macarini, W.; Pimentel, R. L.; Muller, M.; Vassallo, P. F.; Campos, L. C. G.; dos Santos, L.; Luiz, W. B.; Mill, J. G.; Barauna, V. G.; de Carvalho, L. F. das C. e. S. Rapid Diagnosis of COVID-19

- Using FT-IR ATR Spectroscopy and Machine Learning. *Sci Rep* **2021**, *11* (1). <https://doi.org/10.1038/s41598-021-93511-2>.
- (15) Martinez-Cuazitl, A.; Vazquez-Zapien, G. J.; Sanchez-Brito, M.; Limon-Pacheco, J. H.; Guerrero-Ruiz, M.; Garibay-Gonzalez, F.; Delgado-Macuil, R. J.; de Jesus, M. G. G.; Corona-Perezgrovas, M. A.; Pereyra-Talamantes, A.; Mata-Miranda, M. M. ATR-FTIR Spectrum Analysis of Saliva Samples from COVID-19 Positive Patients. *Sci Rep* **2021**, *11* (1). <https://doi.org/10.1038/s41598-021-99529-w>.
 - (16) Nascimento, M. H. C.; Marcarini, W. D.; Folli, G. S.; da Silva Filho, W. G.; Barbosa, L. L.; Paulo, E. H. de; Vassallo, P. F.; Mill, J. G.; Barauna, V. G.; Martin, F. L.; de Castro, E. V. R.; Romão, W.; Filgueiras, P. R. Noninvasive Diagnostic for COVID-19 from Saliva Biofluid via FTIR Spectroscopy and Multivariate Analysis. *Anal Chem* **2022**, *94* (5), 2425–2433. <https://doi.org/10.1021/acs.analchem.1c04162>.
 - (17) Bandeira, C. C. S.; Madureira, K. C. R.; Rossi, M. B.; Gallo, J. F.; da Silva, A. P. M. A.; Torres, V. L.; de Lima, V. A.; Júnior, N. K.; Almeida, J. D.; Zerbini, R. M.; Braz-Silva, P. H.; Lindoso, J. A. L.; da Silva Martinho, H. Micro-Fourier-Transform Infrared Reflectance Spectroscopy as Tool for Probing IgG Glycosylation in COVID-19 Patients. *Sci Rep* **2022**, *12* (1). <https://doi.org/10.1038/s41598-022-08156-6>.
 - (18) Dogan, A.; Gurbanov, R.; Severcan, M.; Severcan, F. CoronaVac (Sinovac) COVID-19 Vaccine-Induced Molecular Changes in Healthy Human Serum by Infrared Spectroscopy Coupled with Chemometrics. *Turkish Journal of Biology* **2021**, *45* (Special Issue 1), 549–558. <https://doi.org/10.3906/biy-2105-65>.
 - (19) Zhang, Z. M.; Chen, S.; Liang, Y. Z. Baseline Correction Using Adaptive Iteratively Reweighted Penalized Least Squares. *Analyst* **2010**, *135* (5), 1138–1146. <https://doi.org/10.1039/b922045c>.
 - (20) Lovatti, B. P. O.; Nascimento, M. H. C.; Neto, Á. C.; Castro, E. V. R.; Filgueiras, P. R. Use of Random Forest in the Identification of Important Variables. *Microchem. J.* **2019**, *145* (December 2018), 1129–1134. <https://doi.org/10.1016/j.microc.2018.12.028>.
 - (21) Oliveira, B. P. L.; Nascimento, M. H. C.; Rainha, K. P.; Alvaro, E. C. S. O.; Filgueiras, P. R. Different Strategies for the Use of Random Forest in NMR Spectra. *J. Chemometrics* **2020**, No. Special issue, 1–10. <https://doi.org/10.1002/cem.3231>.
 - (22) Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45* (1), 5–32. <https://doi.org/10.1023/A:101093340>.
 - (23) Afanador, N. L.; Smolinska, A.; Tran, T. N.; Blanchet, L. Unsupervised Random Forest: A Tutorial with Case Studies. *J. Chemometrics* **2016**, *30* (5), 232–241. <https://doi.org/10.1002/cem.2790>.

- (24) Kennard, R. W.; Stone, L. A. Computer Aided Design of Experiments. *Tecnometrics* **1969**, *11* (1), 137–148.
- (25) Comnea-Stancu, I. R.; Wieland, K.; Ramer, G.; Schwaighofer, A.; Lendl, B. On the Identification of Rayon/Viscose as a Major Fraction of Microplastics in the Marine Environment: Discrimination between Natural and Manmade Cellulosic Fibers Using Fourier Transform Infrared Spectroscopy. *Appl Spectrosc* **2017**, *71* (5), 939–950.
<https://doi.org/10.1177/0003702816660725>.
- (26) Savitzky, A.; E, M. J. *Smoothing and Differentiation of Data by Simplified Least Squares Procedures*; 1951; Vol. 40.
<https://pubs.acs.org/sharingguidelines>.
- (27) Wood, B. R.; Kochan, K.; Bedolla, D. E.; Salazar-Quiroz, N.; Grimley, S. L.; Perez-Guaita, D.; Baker, M. J.; Vongsvivut, J.; Tobin, M. J.; Bambery, K. R.; Christensen, D.; Pasricha, S.; Eden, A. K.; Mclean, A.; Roy, S.; Roberts, J. A.; Druce, J.; Williamson, D. A.; McAuley, J.; Catton, M.; Purcell, D. F. J.; Godfrey, D. I.; Heraud, P. Infrared Based Saliva Screening Test for COVID-19. *Angewandte Chemie - International Edition* **2021**, *60* (31), 17102–17107.
<https://doi.org/10.1002/anie.202104453>.
- (28) Nogueira, M. S.; Leal, L. B.; Macarini, W.; Pimentel, R. L.; Muller, M.; Vassallo, P. F.; Campos, L. C. G.; dos Santos, L.; Luiz, W. B.; Mill, J. G.; Barauna, V. G.; de Carvalho, L. F. das C. e. S. Rapid Diagnosis of COVID-19 Using FT-IR ATR Spectroscopy and Machine Learning. *Sci Rep* **2021**, *11* (1).
<https://doi.org/10.1038/s41598-021-93511-2>.
- (29) Heino, H.; Rieppo, L.; Männistö, T.; Sillanpää, M. J.; Mäntynen, V.; Saarakkala, S. Diagnostic Performance of Attenuated Total Reflection Fourier-Transform Infrared Spectroscopy for Detecting COVID-19 from Routine Nasopharyngeal Swab Samples. *Sci Rep* **2022**, *12* (1), 20358.
<https://doi.org/10.1038/s41598-022-24751-z>.
- (30) Wyllie AL; Fournier J; Casanovas-Massana A; Melissa Campbell; Maria Tokuyama; Pavithra Vijayakumar; Joshua L. Warren; Bertie Geng; M. Catherine Muenker; Adam J. Moore; Chantal B. F. Vogels; Mary E. Petrone; Isabel M. Ott; Peiwen Lu; Arvind Venkataramam; Alice Lu-Culligan; Jonathan Klein; Rebecca Earnest. Saliva or Nasopharyngeal Swab Specimens for Detection of SARS-CoV-2. *New engl. J. Med.* **2020**, *383*, 1283–1286.
<https://doi.org/10.1056/NEJMc2016359>.
- (31) Guleken, Z.; Jakubczyk, P.; Wiesław, P.; Krzysztof, P.; Bulut, H.; Öten, E.; Depciuch, J.; Tarhan, N. Characterization of Covid-19 Infected Pregnant Women Sera Using Laboratory Indexes, Vibrational Spectroscopy, and Machine Learning Classifications. *Talanta* **2022**, *237*.
<https://doi.org/10.1016/j.talanta.2021.122916>.

CAPÍTULO 6

CONCLUSÕES DA TESE



6. CONCLUSÕES DA TESE

Neste trabalho apresentamos diversas abordagens de aprendizagem de máquinas, principalmente o método floresta aleatória, na solução de problemas do ponto de vista da Química. Além disso, ao descrever o princípio histórico da Quimiometria, contribuimos para uma maior compreensão de como os avanços tecnológicos nos equipamentos, a união desses equipamentos a computadores, a popularização e evolução dos microcomputadores, propulsionou o desenvolvimento desta área da Química. A Quimiometria é um campo com ampla aplicação em diferentes áreas da Química. Este campo se desenvolve muito rápido, tornando-se um desafio acompanharmos as diversas abordagens que vão surgindo.

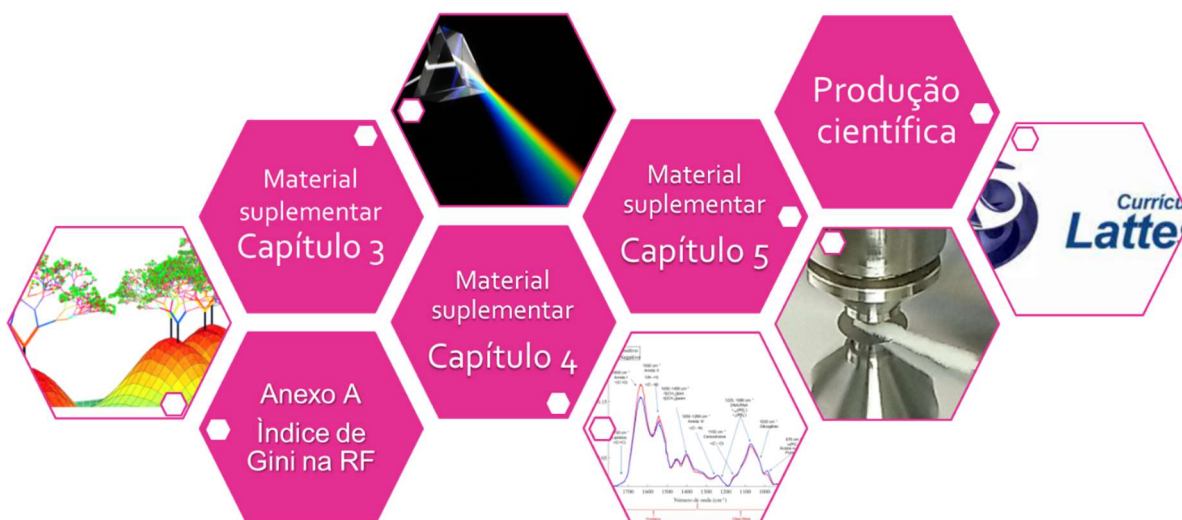
Como foi demonstrado ao longo desta tese, floresta aleatória é um método versátil e eficiente, cuja potencialidade ainda não é bem explorada nas abordagens com dados espectroscópicos na Química Analítica. É possível a utilização da RF não supervisionada para identificar estrutura existente nos dados espectroscópicos como ponto de partida em uma análise de reconhecimento de padrões. Além disso, este modelo não supervisionado pode ser utilizado para selecionar variáveis e até mesmo classificar, mesmo que se esteja diante de amostras não rotuladas, ou seja, sem o conhecimento das classes reais. Essa abordagem tem um alto potencial de aplicação nas investigações iniciais e triagens, podendo ser adaptado às características de cada estudo.

Este trabalho adaptou-se à necessidade do contexto, inserindo-se na busca por contribuir para uma maior compreensão e métodos de diagnóstico alternativos, frente à pandemia causada pelo vírus SARS-CoV-2. Diante deste desafio, é cada vez mais importante projetos de integração entre as áreas de Ciências da Saúde e Ciências Exatas na busca por métodos de diagnósticos e biópsias líquidas, por meio de espectroscopia na região do infravermelho a partir de biofluidos, contribuindo para o estudo de diversas doenças.

Dessa forma, é importante destacar a necessidade e a tendência na Quimiometria de se desenvolver, adaptar, ou aplicar os métodos de aprendizagem de máquina de uma forma heurística. Ou seja, não há uma forma única e padronizada, mas um desafio constante de focarmos nos problemas a serem resolvidos e nas habilidades que precisamos desenvolver para solucioná-los.

CAPÍTULO 7

MATERIAIS SUPLEMENTARES E ANEXOS



- Neste capítulo são apresentados os materiais suplementares dos artigos científicos apresentados nos capítulos 3, 4 e 5 desta tese. Nestes materiais são descritas informações suplementares sobre a metodologia, além de figuras e tabelas complementares citadas nos respectivos artigos.
- O Anexo A apresenta um exemplo simples e didático do cálculo do índice de Gini em uma árvore de classificação.
- Toda a produção científica durante o período do Doutorado em Química também é descrita neste capítulo.

7. MATERIAL SUPLEMENTAR E ANEXOS

7.1. Material Suplementar - O uso de Floresta Aleatória Não Supervisionada na discriminação do óleo diesel: um estudo de caso de adulteração de combustível (Capítulo 3)

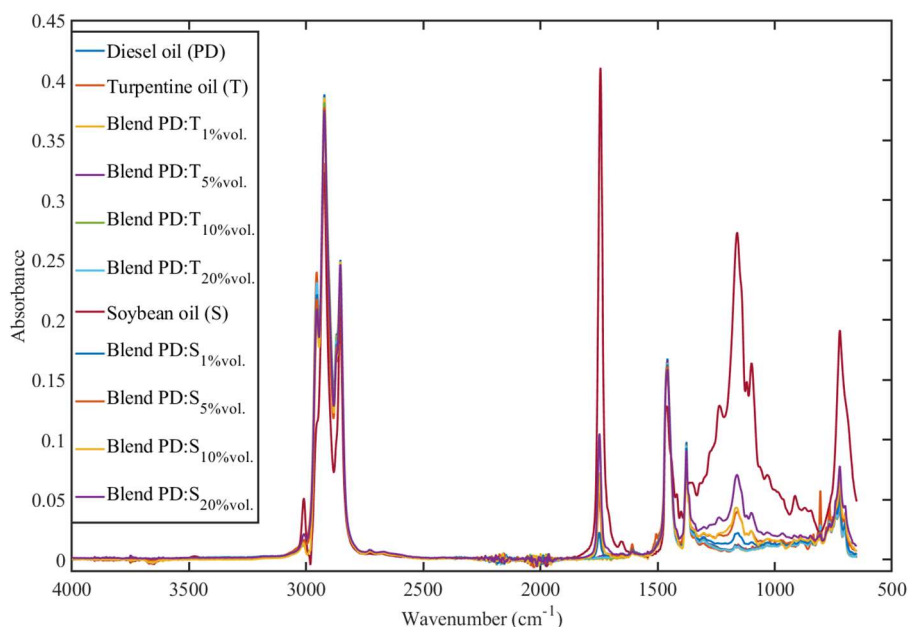


Figura S01. Espectros FT-MIR de óleo diesel (PD) e diesel dopado com adulterantes de terebintina (T) e óleo de soja (S), em nível de concentração de 1 a 20 % do volume.

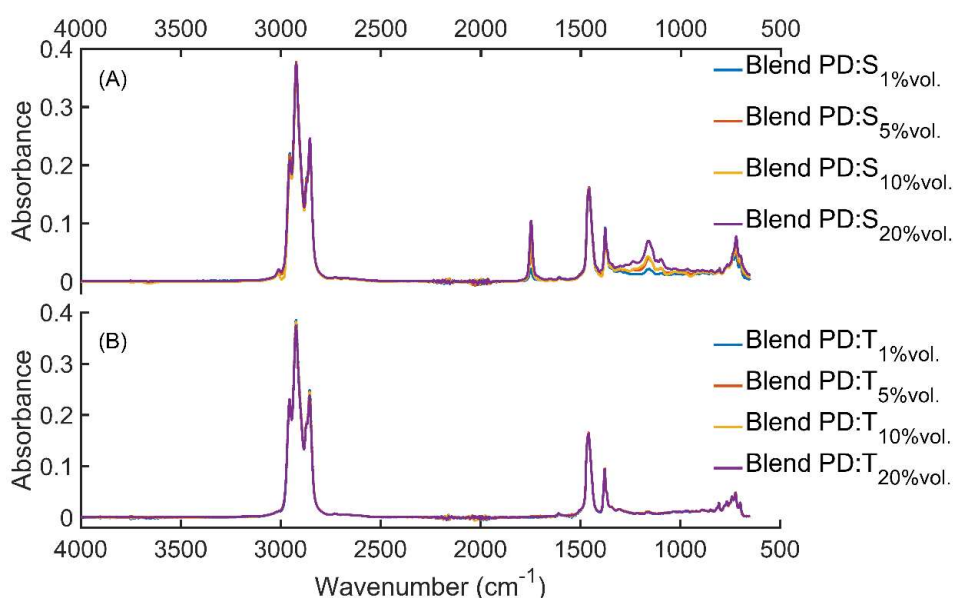


Figura S02. Espectros FT-MIR de curvas padrão de diesel (PD) adulteradas com 1,0; 5,0; 10 e 20% de: (A) óleo de soja (S); (B) terebintina (T)

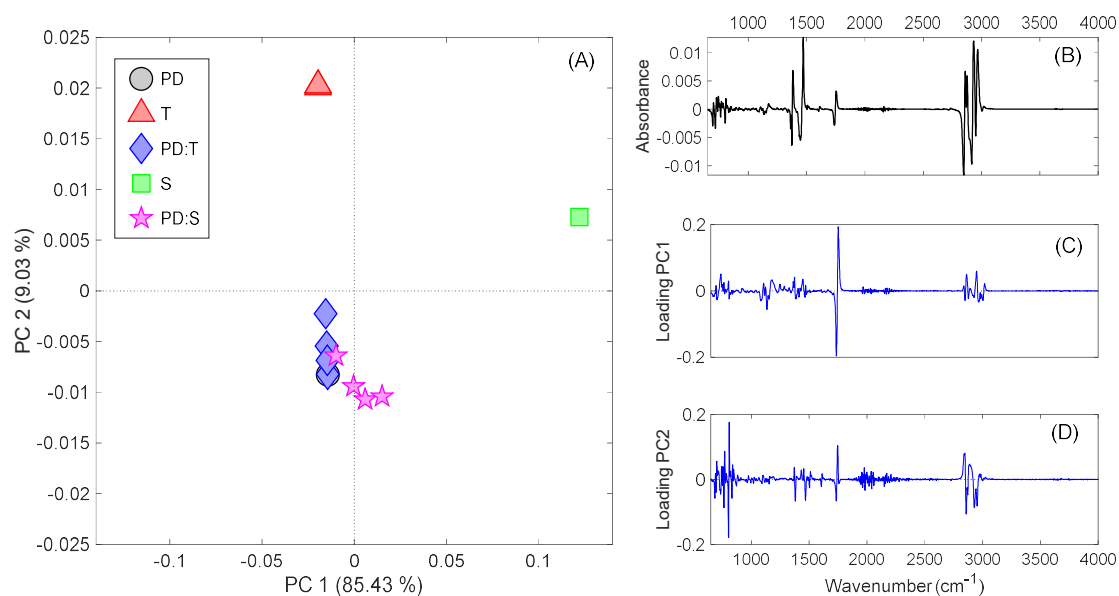


Figura S03. Modelo PCA com diesel (PD) e misturas PD:T e PD:S, nos níveis de concentração 1, 5, 10 e 20 % v/v: (A) Escores PC1 x PC2; (B) primeira derivada da média dos espectros; (C) carregamentos em PC1; (D) cargas em PC2.

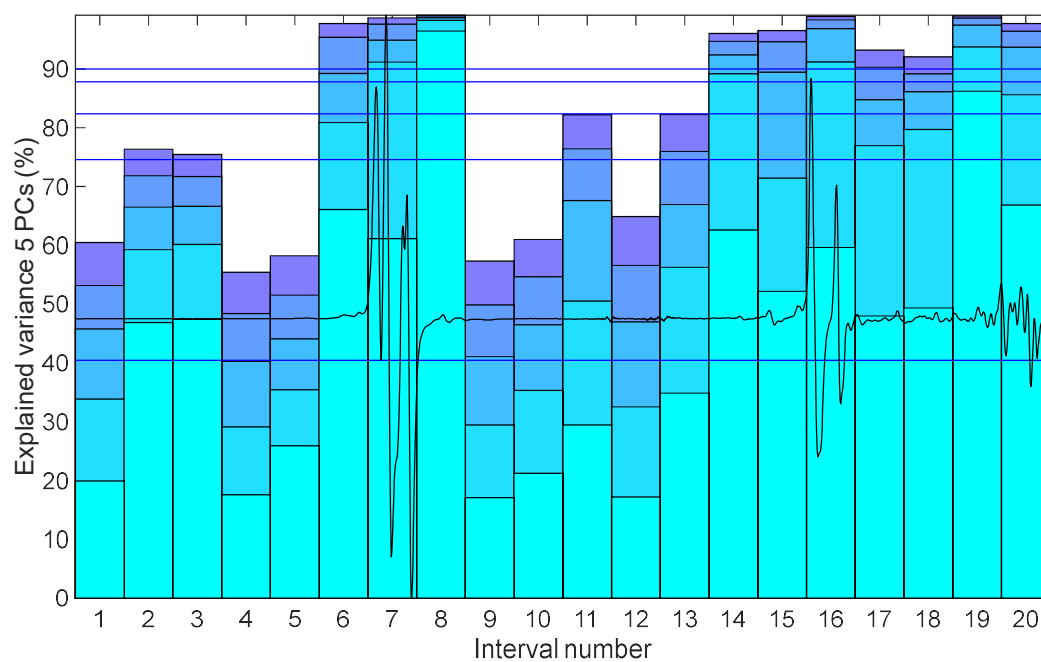


Figura S04. Gráfico de intervalos dos espectros MIR de amostras de diesel (modelo iPCA).

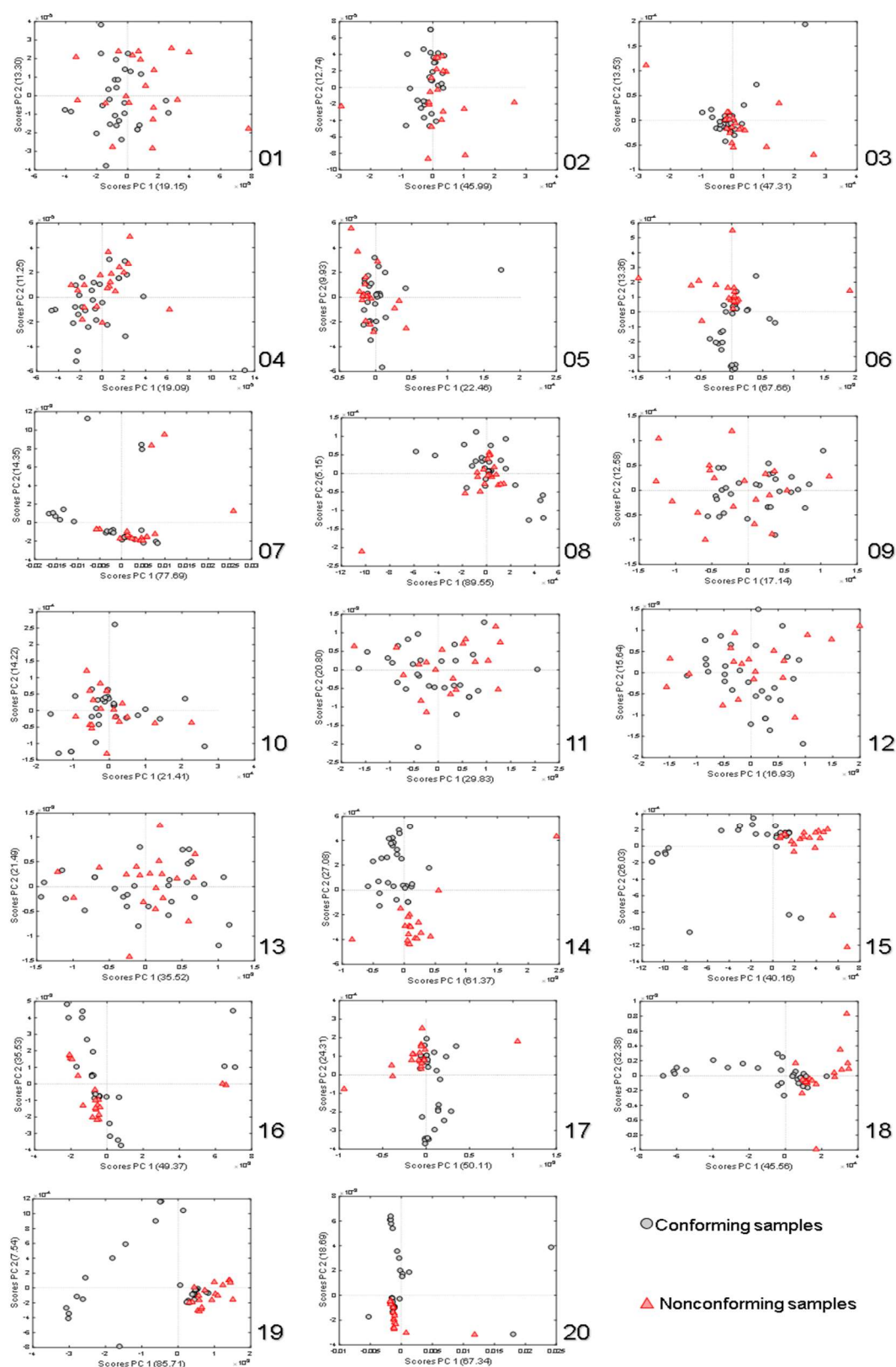


Figura S05. Gráfico de escores de PCA (PC1 versus PC2) de 48 amostras comerciais de óleo diesel, a partir de dados MIR, para cada intervalo de número de onda gerado pelo iPCA. Pré-processamento: primeira derivada.

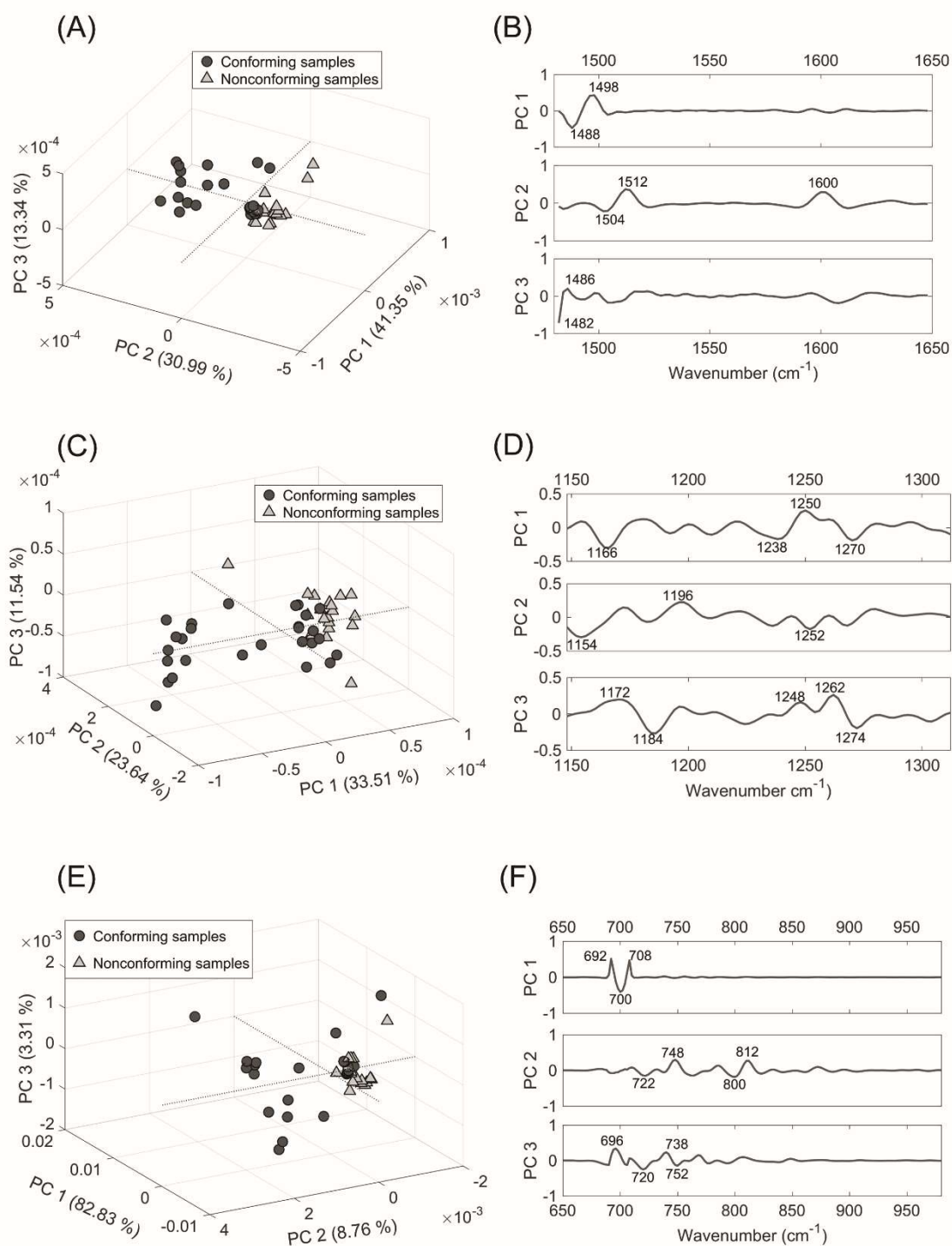


Figura S06. Gráficos de escores e *loadings* de modelos de PCA com intervalos selecionado de variáveis: A e B – escores e *loadings* de modelo de PCA da região espectral MIR de 1500 a 1650 cm^{-1} ; C e D – escores e *loadings* de modelo de PCA da região espectral MIR de 1150 a 1300 cm^{-1} ; E e F – escores e *loadings* de modelo PCA da região espectral MIR de 650 a 950 cm^{-1} .

Tabela S1. Matriz de confusão de modelo SIMCA com parâmetros de desempenho do modelo de região espectral de 1816–1650 cm^{-1} : A) Conjunto de treinamento; B) Conjunto de teste.

A) Treinamento

Matriz de confusão			Parâmetros de desempenho						
Previsto	Real								
	Classe alvo	Classe não-alvo							
Classe alvo	15	1	2	1,00	0,92	0,96	0,06	0	0,94
Classe não-alvo	0	11							

B) Teste externo

Matriz de confusão			Parâmetros de desempenho						
Previsto	Real								
	Classe alvo	Classe não-alvo							
Classe alvo	11	0	2	0,78	1,00	0,86	0	0,21	1,0
Classe não-alvo	3	7							

Tabela S2. Matriz de confusão de modelo SIMCA com parâmetros de desempenho do modelo de região espectral de 1650–1500 cm⁻¹: A) conjunto de treinamento; B) conjunto de teste externo.

A) Treinamento									
Matriz de confusão			Parâmetros de performance						
Previsto	Real								
	Classe alvo	Classe não alvo	PC	Sens	Espec	Exat	TFP	TFN	Prec
	Classe alvo	Classe não alvo							
Classe alvo	13	2	2	0,87	0,83	0,85	0,16	0,13	0,85
Classe não alvo	2	10							
B) Teste									
Matriz de confusão			Parâmetros de performance						
Previsto	Real								
	Classe alvo	Classe não alvo	PC	Sens	Espec	Exat	TFP	TFN	Prec
	Classe alvo	Classe não alvo							
Classe alvo	10	0	2	0,71	1,0	0,81	0	0,28	1,0
Classe não alvo	4	7							

Tabela S3. Matriz de confusão de modelo SIMCA com parâmetros de desempenho do modelo de região espectral de 1300–1150 cm⁻¹: A) conjunto de treinamento; B) conjunto de teste externo.

A) Treinamento									
Matriz de confusão			Parâmetros de performance						
Previsto	Real		PC	Sens	Espec	Exat	TFP	TFN	Prec
	Classe alvo	Classe não alvo							
Classe alvo	13	4	3	0,87	0,67	0,78	0,33	0,13	0,76
Classe não alvo	2	8							

B) Teste									
Matriz de confusão			Parâmetros de performance						
Previsto	Real		PC	Sens	Espec	Exat	TFP	TFN	Prec
	Classe alvo	Classe não alvo							
Classe alvo	9	1	3	0,64	0,86	0,71	0,14	0,35	0,90
Classe não alvo	5	6							

Tabela S4. Matriz de confusão de modelo SIMCA com parâmetros de desempenho do modelo de região espectral de 950–650 cm⁻¹: A) conjunto de treinamento; B) conjunto de teste externo.

A) Treinamento									
Matriz de confusão			Parâmetros de performance						
Previsto	Real		PC	Sens	Espec	Exat	TFP	TFN	Prec
	Classe alvo	Classe não alvo							
Classe alvo	10	5	3	0,67	0,58	0,63	0,41	0,33	0,67
Classe não alvo	5	7							

B) Teste									
Matriz de confusão			Parâmetros de performance						
Previsto	Real		PC	Sens	Espec	Exat	TFP	TFN	Prec
	Classe alvo	Classe não alvo							
Classe alvo	11	4	3	0,79	0,43	0,67	0,57	0,21	0,73
Classe não alvo	3	3							

Referências

- (1) Brasil. Agência Nacional de Petróleo Gás Natural e Biocombustíveis. *Resolução ANP Nº 52 (RANP 52-2010)*; 2010.
- (2) Brasil. Agência Nacional de Petróleo Gás Natural e Biocombustíveis. *Resolução ANP Nº 50, DE 23.12.2013. Agência Nacional do Petróleo, Gás Natural e Biocombustíveis 2013*, 11–13.

7.2. Material suplementar – Diagnóstico não invasivo para COVID-19 a partir de biofluido saliva via espectroscopia FTIR e análise multivariada (Cap. 04)

Informações de suporte

- Tabela S-1. Previsão de classe do conjunto de dados de teste de quatro modelos de classificação.
- Figura S01. Histograma da idade (anos) dos pacientes. (A) Diagnóstico RT-qPCR positivo; (B) Diagnóstico RT-qPCR negativo.
- Figura S02. Histograma de genes-alvo de análise RT-PCR. (A) Gene N; (B) Gene ORF1ab.
- Figura S03. Conjunto de dados espectrais MIR de amostras de saliva de n=237 pacientes com diagnóstico de RT-qPCR para infecção por COVID-19. A) positivo (n=138 amostras) e B) negativo (n=99 amostras).
- Figura S04. Conjunto de dados de espectros MIR originais de amostras de saliva (n=265) com o espectro médio (linha vermelha) e com diferentes abordagens de correção de linha de base avaliadas: A) espectros brutos B) Espectros após a linha de base corrigidos por meio do algoritmo airPLS; C) Espectros após o pré-processamento de primeira derivada; D) Espectros após pré-processamento de SNV e suavização de Savitzky-Golay.
- Figura S05. Espectro derivado médio de amostras positivas (azul) e negativas (vermelho) com identificação da maior frequência no modelo URF (linha pontilhada).
- Figura S06. MIR espectros truncados de 237 amostras (A) com um espectro médio (linha vermelha), e MIR espectros processados via segunda derivada com um espectro médio (linha vermelha).
- Figura S07. Classes reais e consensuais previstas para n = 165 amostras de saliva no conjunto de dados de treinamento (A) e para 72 amostras de saliva no conjunto de dados de teste (B).
- Tabela S-2. Matriz de confusão de modelos de classificação do conjunto de dados de treinamento e teste.
- Tabela S-3. Tabela cruzada de sintomas de teste χ^2 (de acordo com os Centros de Controle e Prevenção de Doenças, CDC) do conjunto de dados de coortes participantes.
- Figura S08. Boxplot dos dias dos primeiros sintomas do conjunto de dados (n=237).
- Tabela S-4. Parâmetros de qualidade de previsão de novo conjunto de dados (n=59) a partir de modelos aplicados.
- Referências Complementares

Tabela S-1. Classes previstas para conjunto de teste a partir de 4 modelos de classificação.

Samples	Classe real	Classe prevista pelos modelos				
		GA-LDA*	SPA-LDA	PLS-DA*	PSO-PLS-DA*	*Consensus class
1	1	1	1	2	2	1
2	1	1	1	1	1	1
3	1	2	2	2	2	2
4	1	1	1	1	1	1
5	1	1	1	2	1	1
6	1	1	1	2	2	2
7	1	1	1	1	1	1
8	1	1	1	1	1	1
9	1	1	1	1	1	1
10	1	1	1	2	1	1
11	1	1	1	1	1	1
12	1	1	1	1	1	1
13	1	1	1	1	1	1
14	1	1	1	1	1	1
15	1	1	1	1	1	1
16	1	1	1	2	1	1
17	1	1	2	2	1	1
18	1	1	1	1	1	1
19	1	1	1	1	1	1
20	1	1	1	1	1	1
21	1	1	1	2	1	1
22	1	1	1	2	2	1
23	1	1	1	1	1	1
24	1	1	1	1	1	1
25	1	1	1	1	2	1
26	1	1	1	1	1	1
27	1	1	1	1	1	1
28	1	1	1	1	1	1
29	1	1	1	1	1	1
30	1	1	1	1	1	1
31	1	1	1	1	1	1
32	1	1	1	1	1	1
33	1	2	1	2	2	2
34	1	1	1	1	2	1
35	1	1	1	1	1	1
36	1	1	1	1	1	1
37	1	1	1	1	1	1
38	1	1	1	1	1	1
39	1	1	1	1	1	1

Continua...

Classe prevista pelos modelos						
Samples	Classe real	GA-LDA*	SPA-LDA	PLS-DA*	PSO-PLS-DA*	Consensus class
Continuação						
40	1	1	1	1	1	1
41	1	1	1	1	1	1
42	2	2	1	2	2	2
43	2	2	2	2	2	2
44	2	2	1	2	2	2
45	2	2	2	2	2	2
46	2	2	2	2	2	2
47	2	2	2	2	2	2
48	2	1	2	1	1	1
49	2	1	1	1	1	1
50	2	2	2	2	2	2
51	2	2	2	2	2	2
52	2	2	1	2	2	2
53	2	1	1	1	1	1
54	2	1	2	2	2	2
55	2	1	1	2	1	1
56	2	2	1	1	2	2
57	2	2	1	2	2	2
58	2	2	1	2	2	2
59	2	2	1	2	2	2
60	2	2	2	2	2	2
61	2	2	1	1	1	1
62	2	2	2	2	2	2
63	2	1	1	1	1	1
64	2	2	1	2	2	2
65	2	2	1	2	2	2
66	2	2	2	2	2	2
67	2	1	1	1	1	1
68	2	2	2	2	2	2
69	2	2	2	2	2	2
70	2	1	1	1	1	1
71	2	2	2	2	2	2
72	2	1	2	2	2	2

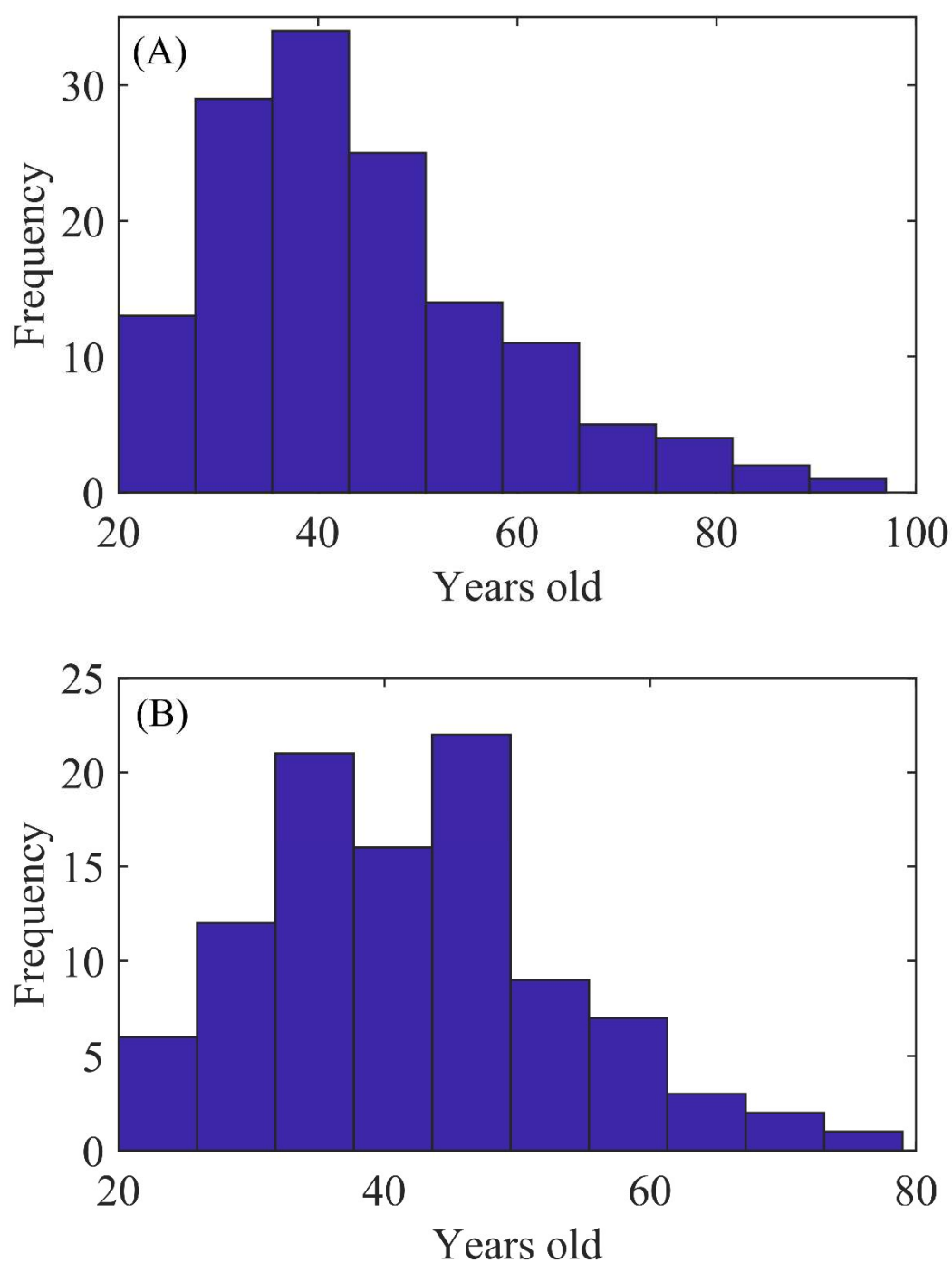


Figura S01. Histograma da idade (anos) do paciente. (A) Diagnóstico RT-qPCR positivo; (B) Diagnóstico RT-qPCR negativo.

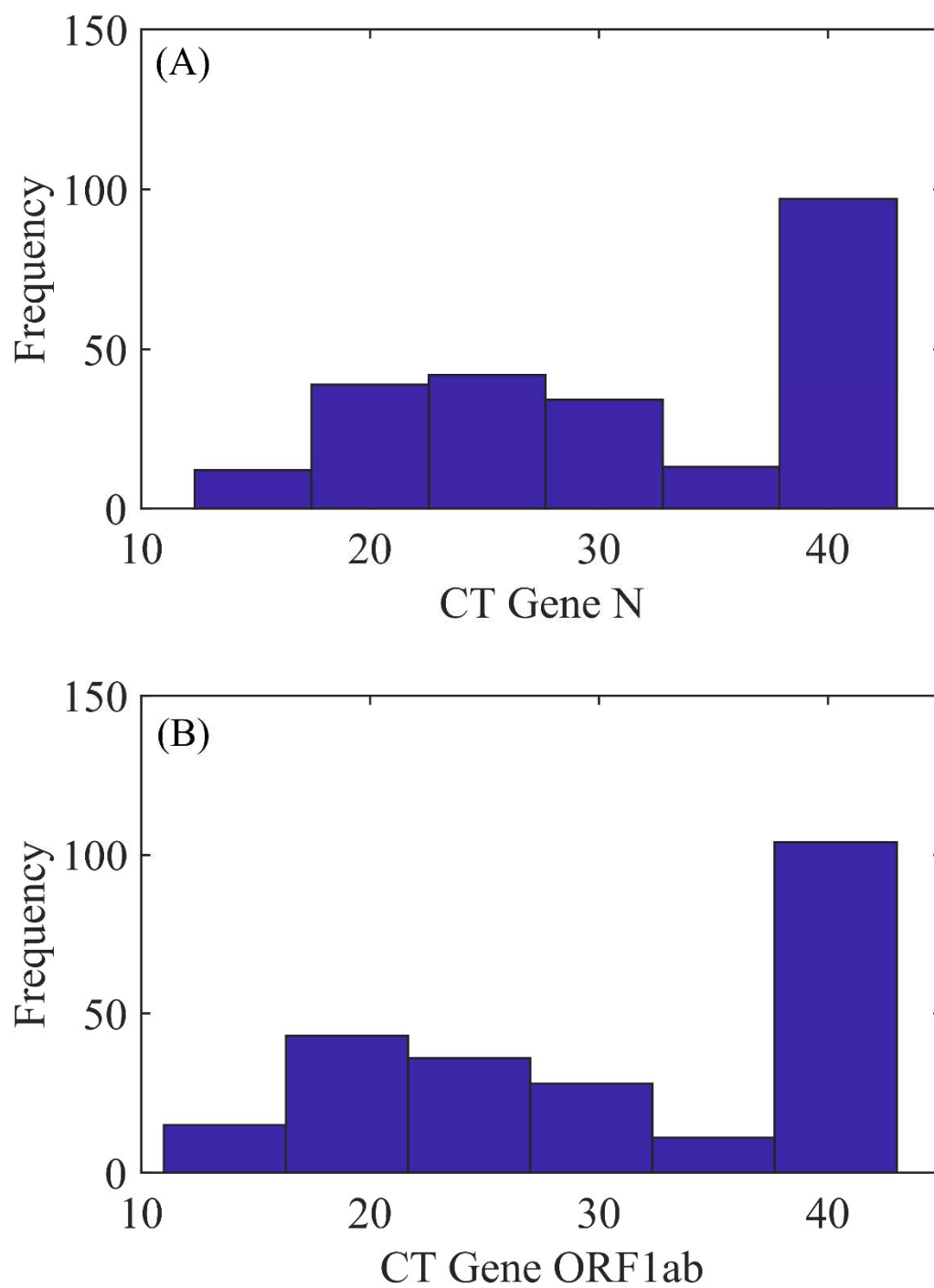


Figura S02. Histograma genes-alvo de análise RT-PCR. (A) Gene N; (B) Gene ORF1ab.

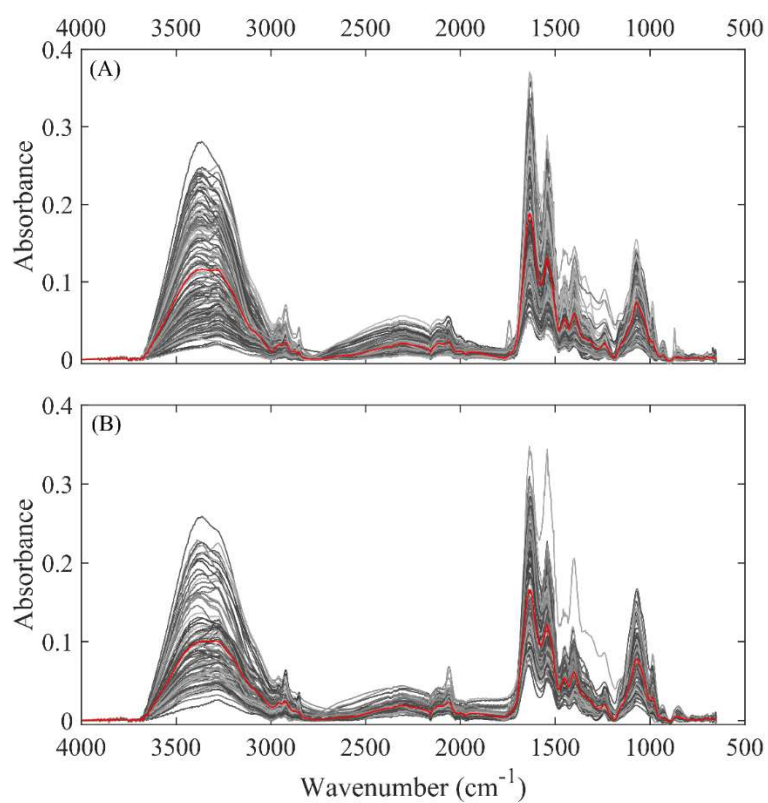


Figura S03. Conjunto de dados espectrais MIR de amostras de saliva de $n=237$ pacientes com diagnóstico de RT-qPCR para infecção por COVID-19. A) positivo ($n=138$ amostras) e B) negativo ($n=99$ amostras).

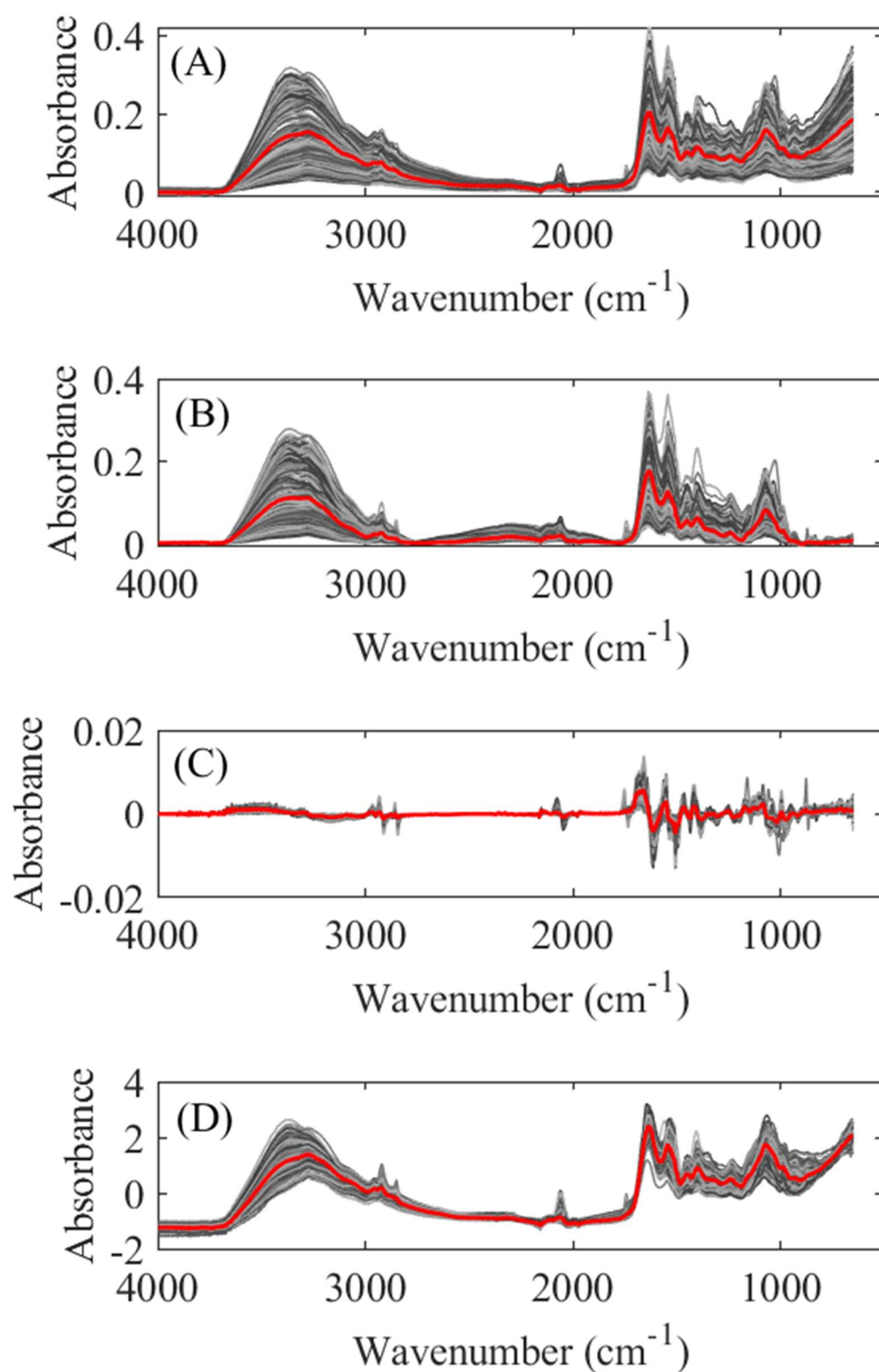


Figura S04. Conjunto de dados de espectros MIR originais de amostras de saliva ($n=265$) com o espectro médio (linha vermelha) e com diferentes abordagens de correção de linha de base avaliadas: A) espectros brutos B) Espectros após a linha de base corrigidos por meio do algoritmo airPLS; C) Espectros após o pré-processamento de primeira derivada aplicado à correção de linha de base D) Espectros após pré-processamento de SNV e suavização de Savitzky-Golay.

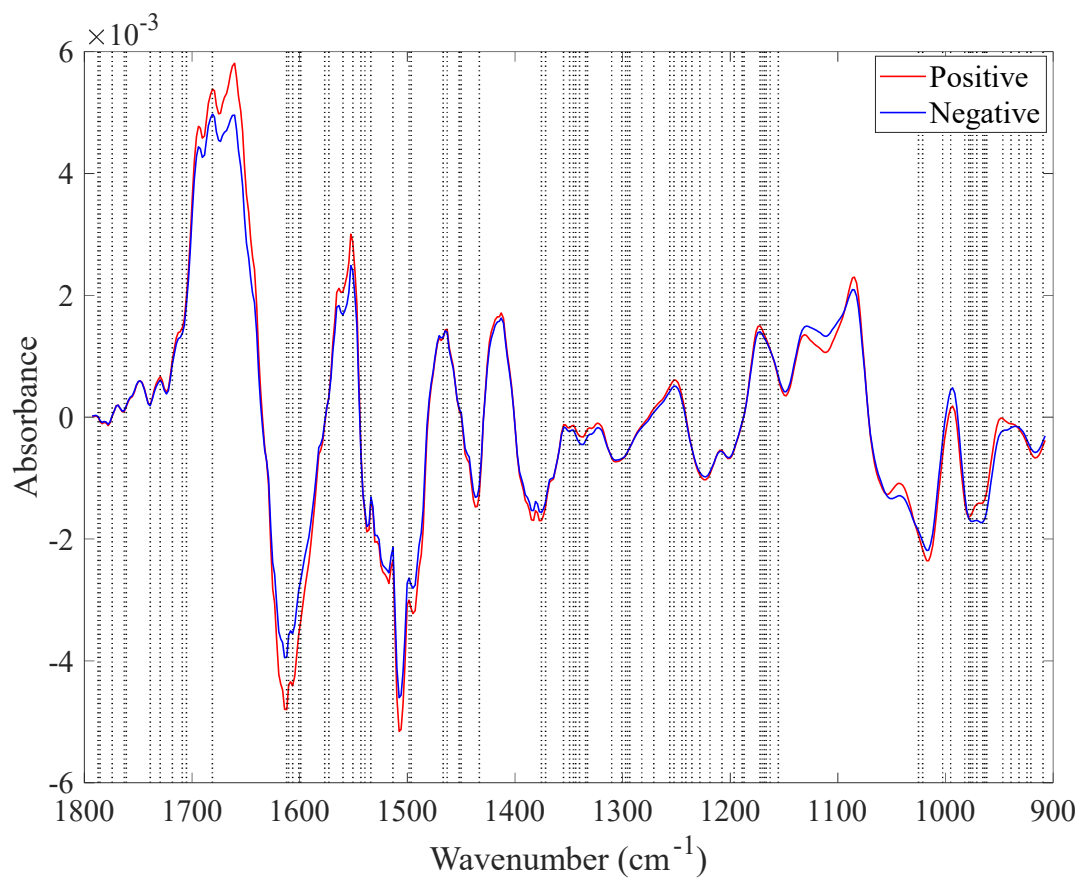


Figura S05. Espectro derivado médio de amostras positivas (vermelho) e negativas (azul) com identificação da maior frequência no modelo URF (linha pontilhada).

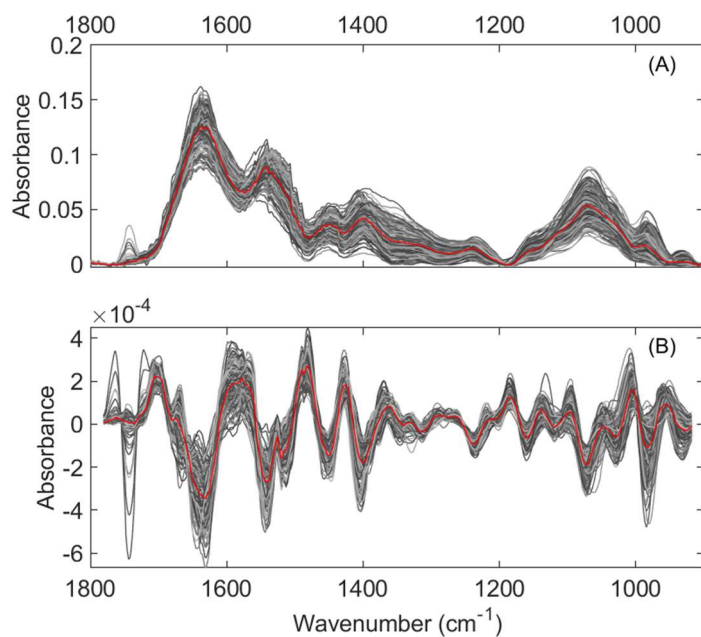


Figura S06. MIR espectros truncados (após correção de linha de base) de 237 amostras (A) com um espectro médio (linha vermelha), e MIR espectros processados via segunda derivada com um espectro médio (linha vermelha).

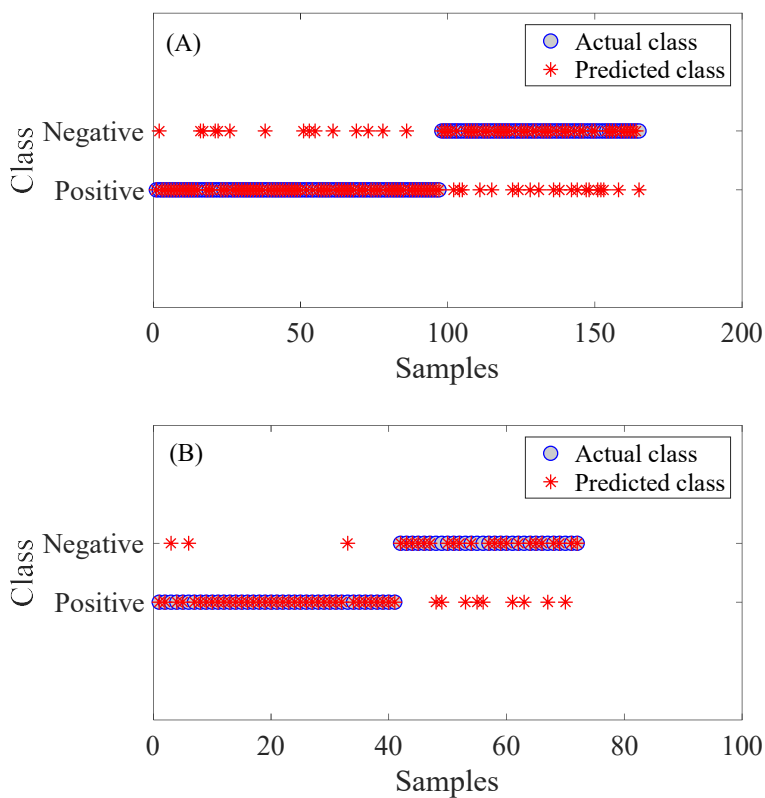


Figura S07. Classes reais e consensuais previstas para $n = 165$ amostras de saliva no conjunto de dados de treinamento (A) e para $n = 72$ amostras de saliva no conjunto de dados de teste (B).

Tabela S-2. Matriz de confusão de modelos de classificação do conjunto de dados de treinamento e teste.

		GA-LDA		SPA-LDA		PLS-DA		PSO-PLSDA	
		Cl.1	Cl.2	Cl.1	Cl.2	Cl.1	Cl.2	Cl.1	Cl.2
VP	Treino	84	46	67	43	68	52	77	52
	Teste	39	22	36	21	31	23	34	23
VN	Treino	46	84	43	67	52	68	52	77
	Teste	22	39	21	36	23	31	23	34
FP	Treino	22	13	25	30	16	29	16	20
	Teste	9	2	10	5	8	10	8	7
FN	Treino	13	22	30	25	29	16	20	16
	Teste	2	9	5	10	10	8	7	8

VP = Verdadeiro positivo; VN = Verdadeiro negativo; FP = Falso positivo; FN = Falso negativo; Cl.1 = Classe 1; Cl.2 = Classe 2.

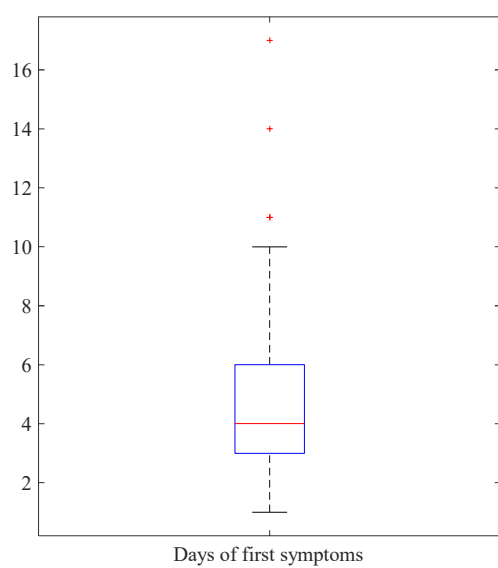


Figura S08. Boxplot dos dias dos primeiros sintomas do conjunto de dados ($n=237$).

Tabela S-4. Parâmetros de qualidade de previsão de novo conjunto de dados ($n=59$).

Modelo	Conjunto de modelos							Parâmetro de qualidade de previsão do novo conjunto de dados										
	SET	TRAIN	OUTLIER	VAL	OUTLIER	VL/PC	VAR	SENS	SPEC	NER	ER	NOT ASSIGNED	VPP	VPN	TFP	TFN	EXAT	MCC
GA-LDA	246	172	7	72	2	0	34	76%	32%	54%	46%	0%	56%	54%	68%	24%	55%	0.08
PLSDA	246	172	7	74	2	7	464	70%	48%	59%	41%	0%	57%	61%	52%	30%	59%	0.18
PSO-PLSDA	246	172	7	74	2	9	45	77%	48%	62%	38%	0%	63%	65%	52%	23%	63%	0.26
Consensus class	----	----	----	----	----	----	----	70%	41%	56%	44%	0%	58%	56%	58%	29%	57%	0.13

7.3. Material suplementar - Floresta Aleatória Não Supervisionada como alternativa para reconhecimento de padrões e diagnóstico de COVID-19 a partir de swab nasofaríngeo e espectroscopia MIR (Cap. 05).

Neste material suplementar estão apresentados:

Figura S01. Representação da etapa de subtração de espectros de swab em conjunto de dados número 03 (*swab* normal com procedimento de secagem).

Tabela S01 Parâmetros de desempenho de modelos de classificação para a classe Positivo, de conjuntos de treinamento e teste (linha sombreada) a partir de variáveis espectrais pré-processadas.

Tabela S02 Parâmetros de desempenho de modelos de classificação para a classe Positivo, de conjuntos de treinamento e teste (linha sombreada) a partir de variáveis espectrais pré-processadas.

Tabela S03. Parâmetros de desempenho de modelos de classificação URF-PCoA-RF, a partir de região de Fingerprint, para a classe Positivo, de conjuntos de treinamento e teste (linha sombreada).

Figura S02 Gráficos de coordenadas principais, a partir de modelo URF-PCoA (distância Euclidiana) a partir de região *Fingerprint* de conjuntos de dados avaliados. A) Conjunto *swab* fino com secagem; B) Conjunto *swab* fino sem secagem; C) Conjunto *swab* normal com secagem; D) Conjunto *swab* normal sem secagem.

Figura S03 Gráficos de coordenadas principais, a partir de modelo URF-PCoA (distância Euclidiana) a partir de região *Fingerprint* de conjuntos de dados avaliados. A) Conjunto *swab* fino + *swab* normal com secagem; B) Conjunto *swab* fino + *swab* normal sem secagem.

Tabela S04 Divisão de conjuntos de dados em treinamento e teste e número de outliers identificados nos modelos de classificação.

Tabela S05 Divisão de conjuntos de dados em treinamento e teste e número de outliers identificados nos modelos de classificação.

Figura S04 Parâmetros otimizados para modelos de classificação RF a partir de variáveis selecionadas pelo método URF (A) e variáveis espectrais na região de *fingerprint* (C).

Figura S05 Parâmetros otimizados para modelos de classificação URF-PCoA-RF a partir de dois métodos de atribuição de classes, para amostras não-rotuladas (A) e rotuladas (B).

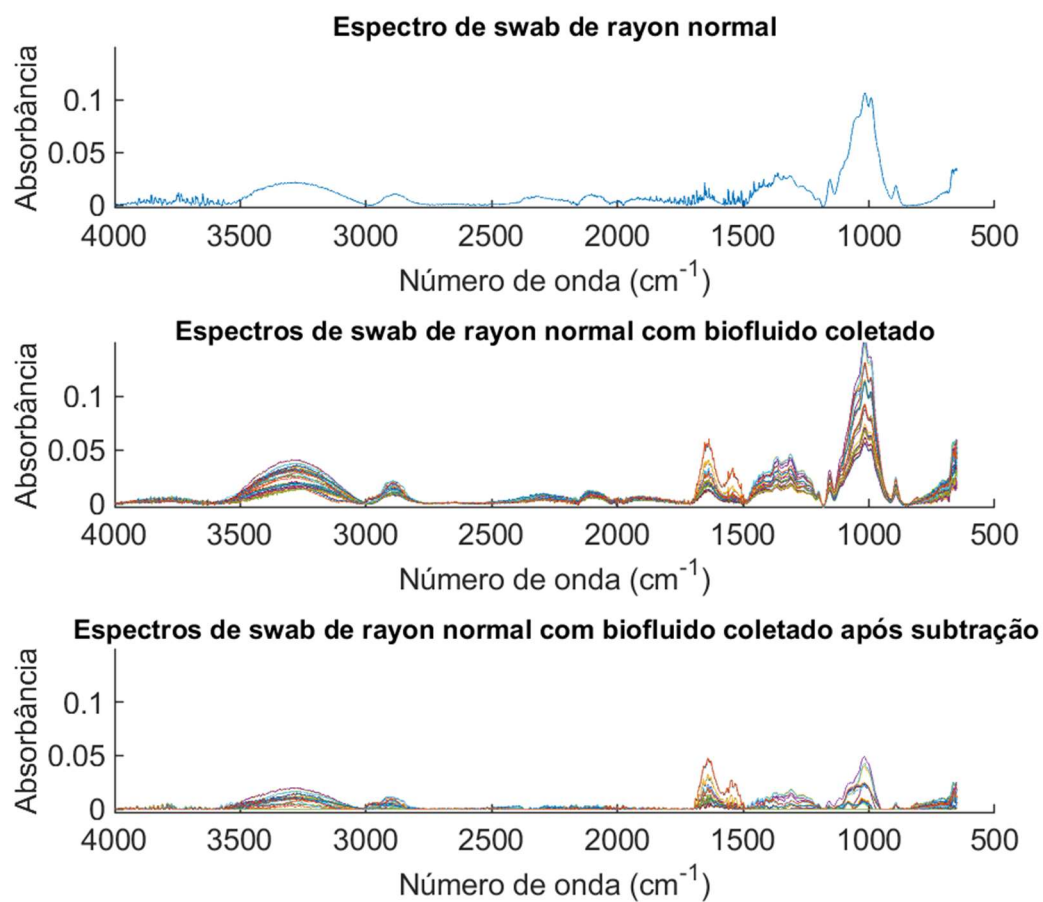


Figura S01. Representação da etapa de subtração de espectros de swab em conjunto de dados número 03 (*swab* normal com procedimento de secagem).

Tabela S01 Parâmetros de desempenho de modelos de classificação para a classe Positivo, de conjuntos de treinamento e teste (linha sombreada) a partir de variáveis espectrais pré-processadas com seleção de variáveis (A), espectros inteiros (B) e região de fingerprint (C).

Conjunto dados	Modelo	Divisão espectro	Pré-processamento	VL	SENS (%)	ESPEC (%)	F-Score	EXAT (%)
01	PLSDA	A	Normalização	6	80,4	90,6	0,841	85,9
					60,0	63,6	0,600	61,9
		B	SNV; 1ª Derivada	8	75,0	84,1	0,796	79,2
					62,5	71,4	0,667	66,7
		C	SNV; 1ª Derivada	10	81,5	84,4	0,838	82,8
					70,0	66,7	0,737	68,7
	GALDA	A	Centrado na média	--	70,0	74,5	0,707	72,4
					76,2	75,0	0,744	75,6
		B	1ª Derivada (15)	--	83,6	78,0	0,821	80,3
					87,5	76,2	0,840	82,2
		C	Centrado na média	--	87,3	82,0	0,857	84,8
					87,5	76,2	0,840	82,2
02	PLSDA	A	Normalização	--	47,4	68,7	0,519	58,7
					50,0	62,5	0,518	56,7
		C	MSC	--	59,6	57,8	0,519	58,7
					71,4	43,7	0,606	56,7
	GALDA	A	2ª derivada	17	78,7	85,7	0,807	82,4
					70,4	60,0	0,655	64,9
		B	1ª derivada (7)	7	77,9	66,7	0,752	72,7
					80,0	59,3	0,653	70,2
		C	1ª derivada (15)	10	76,8	68,8	0,706	73,1
					66,7	51,8	0,635	59,6
03	GALDA	A	MSC	--	77,0	79,7	0,770	78,5
					77,2	76,7	0,764	77,2
		B	2ª derivada (15)	--	89,9	83,9	0,879	87,0
					70,0	85,2	0,764	77,2
		C	SNV; 2ª derivada (15)	--	74,9	80,6	0,806	80,0
					90,0	66,7	0,818	78,9
	RF	A	SNV; 1ª derivada (7)	--	31,7	71,4	0,388	52,6
					33,3	83,3	0,439	59,6
		C	Centrado na média	--	45,8	56,2	0,471	51,3
					50,0	60,0	0,514	55,3
	GALDA	A	MSC	--	100	100	1	100
					92,9	100	0,963	93,5
		B	SNV; 2ª derivada (9)	--	100	100	1	100
					100	80,5	0,714	84,3
		C	2ª derivada (9)	--	100	100	1	100
					100	78,0	0,689	82,3
03	RF	A	Normalização	--	75,0	50,0	0,775	68,2
					40,0	100	0,571	46,0
		C	SNV	--	96,8	33,3	0,923	86,5
					76,7	0	0,793	65,7

A = Variáveis previamente selecionadas pelo URF-PCoA; B = Espectro inteiro (exceto região 2800 – 1800 cm⁻¹); C = Espectro região de *Fingerprint* (1800 - 900 cm⁻¹); SENS = Sensibilidade; ESPEC = Especificidade; EXAT = Exatidão.

Tabela S02 Parâmetros de desempenho de modelos de classificação para a classe Positivo, de conjuntos de treinamento e teste (linha sombreada) a partir de variáveis espectrais pré-processadas com seleção de variáveis (A), espectros inteiros (B) e região de fingerprint (C).

Conjunto dados	Modelo	Divisão espectro	Pré-processamento	VL	SENS (%)	ESPEC (%)	F-Score	EXAT (%)
05	PLSDA	A	SNV; 2 ^a derivada (15)	2	74,1	53,4	0,72	65,7
					67,5	62,9	0,7	65,7
		B	SNV; 2 ^a derivada (15)	8	83,6	78,6	0,836	81,4
					83,3	57,1	0,8	73,7
		C	MSC	10	68,9	80,7	0,750	74,0
					78,6	25,0	0,71	59,1
	GALDA	A	MSC	--	83,1	71,4	0,812	78,1
					70,0	85,2	0,778	76,1
		B	2 ^a derivada (9)	--	84,5	84,1	0,861	84,3
					82,5	74,1	0,825	79,1
		C	MSC	--	79,7	68,4	0,781	74,8
					100	75,0	0,933	90,9
06	RF	A	Centrado na média	--	68,9	54,8	0,689	63,1
					34,6	88,9	0,486	56,8
		C	MSC	--	75,5	46,6	0,711	63,7
					80,8	38,9	0,724	63,6
	PLSDA	A	MSC	12	71,8	81,5	0,771	76,1
					66,0	70,6	0,704	67,9
		B	SNV; 1 ^a derivada (15)	3	58,1	74,7	0,656	65,2
					48,9	85,3	0,613	64,2
		C	2 ^a derivada (9)	3	62,0	71,2	0,670	66,1
					59,3	82,3	0,691	69,1
06	GALDA	A	MSC	--	84,5	77,8	0,836	81,5
					72,3	73,5	0,756	72,8
		B	SNV; 2 ^a derivada (15)	--	77,7	71,2	0,777	74,9
					80,8	67,6	0,792	75,3
		C	SNV; 2 ^a derivada (9)	--	80,2	60,0	0,764	71,6
					78,7	76,5	0,804	77,8
	RF	A	Centrado na média	--	60,0	53,3	0,617	57,1
					77,4	87,0	0,827	81,5
		C	Centrado na média	--	61,6	56,5	0,636	59,4
					74,2	60,9	0,730	68,5

A = Variáveis previamente selecionadas pelo URF-PCoA; B = Espectro inteiro (exceto região 2800 – 1800 cm⁻¹); C = Espectro região de *Fingerprint* (1800 - 900 cm⁻¹); SENS = Sensibilidade; ESPEC = Especificidade; EXAT = Exatidão.

Tabela S03. Parâmetros de desempenho de modelos de classificação URF-PCoA-RF, a partir de região de Fingerprint, para a classe Positivo, de conjuntos de treinamento e teste (linha sombreada).

Conjunto dados	Modelo	Método	Pré-processamento	PCo	SENS (%)	ESPEC (%)	F-Score	EXAT (%)
01	URF/RF	A	----	4	96,5	66,7	0,859	82,9
					100	85,7	0,943	93,5
		B	----	----	94,0	80,4	0,870	86,8
02	URF/RF				57,1	66,7	0,585	62,2
		A	----	2	82,3	85,9	0,829	84,2
					92,6	100	0,961	96,5
03	URF/RF	B	----	----	93,6	94,3	0,936	93,9
					74,1	90,0	0,800	82,5
		A	----	3	92,0	92,0	0,920	92,0
04	URF/RF				100	90,9	0,956	95,4
		B	----	----	----	----	----	----
		A	----	2	75,6	58,3	0,809	71,9
05	URF/RF				100	66,7	0,857	83,3
		B	----	----	----	----	----	----
		A	SNV / 1 ^a DERIVADA (9)	4	90,69	85,9	0,895	88,5
06	URF/RF				91,8	96,7	0,944	94,0
		B		----	97,8	50,0	0,841	78,2
					92,5	22,2	0,755	64,2
06	URF/RF	A	----	2	67,5	77,3	0,740	71,4
					80,9	82,5	0,819	81,7
		B	----	----	99,1	74,1	0,908	88,4
					97,9	79,4	0,920	90,1

A = Método com estimativa de atribuição de classes pelo gráfico PCoA (amostras não rotuladas). B

= Método com atribuição de classe real (amostras rotuladas).

SENS = Sensibilidade; ESPEC = Especificidade; EXAT = Exatidão.

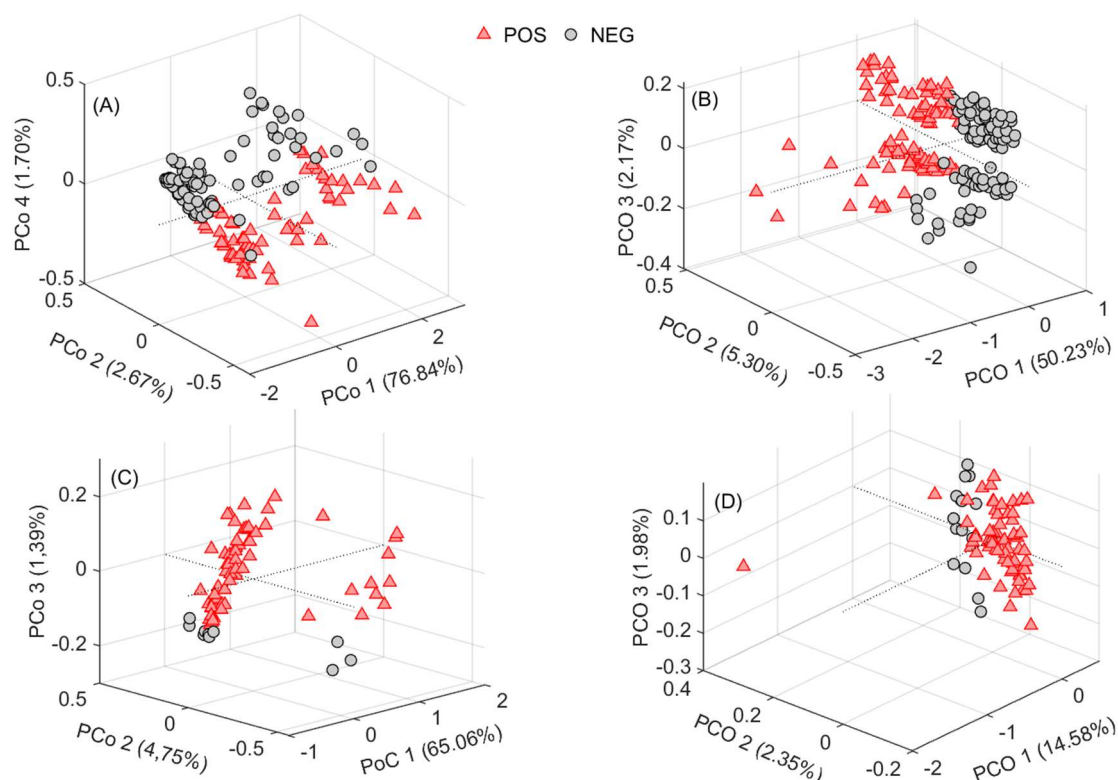


Figura S02 Gráficos de coordenadas principais, a partir de modelo URF-PCoA (distância Euclidiana) a partir de região *Fingerprint* de conjuntos de dados avaliados. A) Conjunto swab fino com secagem; B) Conjunto swab fino sem secagem; C) Conjunto swab normal com secagem; D) Conjunto swab normal sem secagem.

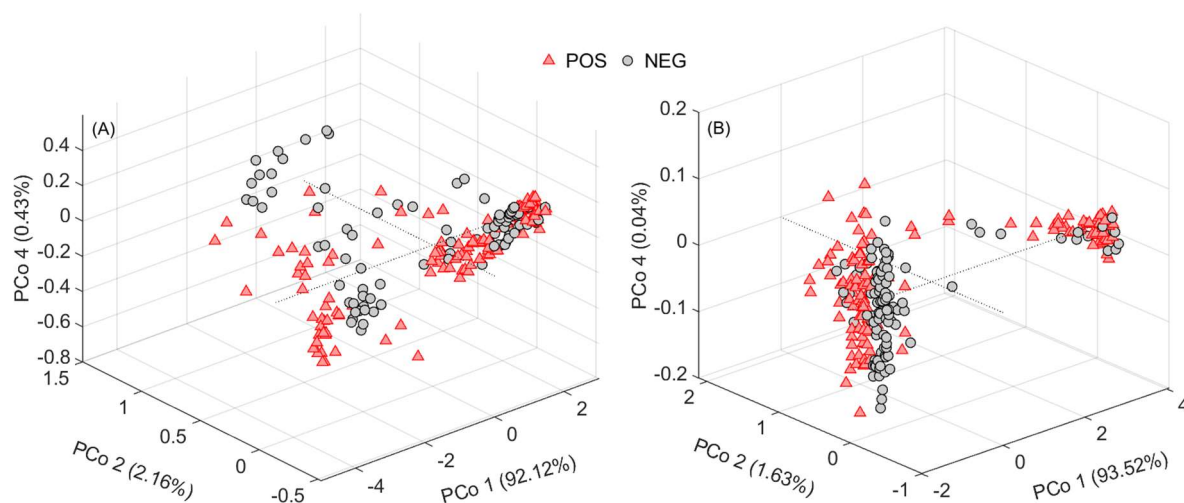


Figura S03 Gráficos de coordenadas principais, a partir de modelo URF-PCoA (distância Euclidiana) a partir de região *Fingerprint* de conjuntos de dados avaliados. A) Conjunto swab fino + swab normal com secagem; B) Conjunto swab fino + swab normal sem secagem.

Tabela S04 Divisão de conjuntos de dados em treinamento e teste e número de outliers identificados nos modelos de classificação.

Conjunto dados	Modelo	Divisão espectro	Treino		Teste		Outliers	
			POS	NEG	POS	NEG	Treino	Teste
01 (n=151) POS=71 NEG=80	PLSDA	A	46	53	11	10	7	24
		B	52	44	8	7	10	30
		C	54	45	6	10	7	29
	GALDA	A	50	55	21	24	1	0
		B	50	55	21	24	1	0
		C	50	55	21	24	1	0
	RF	A	57	64	14	16	0	0
		C	57	64	14	16	0	0
02 (n=190) POS=90 NEG=100	PLSDA	A	61	70	27	30	2	0
		B	60	68	27	30	5	0
		C	61	69	27	30	3	0
	GALDA	A	61	69	27	30	3	0
		B	62	69	27	30	2	0
		C	62	68	27	30	3	0
	RF	A	63	70	27	30	0	0
		C	72	80	18	20	0	0
03 (n=72) POS=61 / NEG=11	GALDA	A	16	6	37	4	0	9
		B	20	6	41	5	0	0
		C	20	6	41	5	0	0
	RF	A	31	6	30	5	0	0
		C	31	6	30	5	0	0
04 (n=81) POS=66 / NEG=15	GALDA	A	20	8	46	7	0	0

A = Variáveis previamente selecionadas pelo URF-PCoA; B = Espectro inteiro (exceto região 2800 – 1800 cm⁻¹); C = Espectro região de *Fingerprint* (1800 - 900 cm⁻¹); SENS = Sensibilidade; ESPEC = Especificidade; EXAT = Exatidão.

Tabela S05 Divisão de conjuntos de dados em treinamento e teste e número de outliers identificados nos modelos de classificação.

Conjunto de dados	Modelos		Treino		Teste		Outliers	
			POS	NEG	POS	NEG	Treino	Teste
05 (n=223) POS=132 NEG=91	PLSDA	A SNV; 2 ^a derivada (15)	85	58	40	27	13	0
		B SNV; 2 ^a derivada (15)	73	56	12	7	27	48
		C MSC	74	57	14	8	25	45
	GALDA	A MSC	83	63	40	27	10	0
		B 2 ^a derivada (9)	84	63	40	27	9	0
		C MSC	74	57	14	8	25	45
	RF	A Centrado na média	106	73	26	18	0	0
		C MSC	106	73	26	18	0	0
06 (n=271) POS=156 NEG=115	PLSDA	A MSC	103	81	47	34	6	0
		B SNV; 1 ^a derivada (15)	105	79	47	34	6	0
		C 2 ^a derivada (9)	100	80	47	34	10	0
	GALDA	A MSC	103	81	47	34	6	0
		B SNV; 2 ^a derivada (15)	103	80	47	34	7	0
		C SNV; 2 ^a derivada (9)	101	75	47	34	14	0
	RF	A Centrado na média	125	92	31	23	0	0
		C Centrado na média	125	92	31	23	0	0

A = Variáveis previamente selecionadas pelo URF-PCoA; B = Espectro inteiro (exceto região 2800 – 1800 cm⁻¹); C = Espectro região de *Fingerprint* (1800 - 900 cm⁻¹); SENS = Sensibilidade; ESPEC = Especificidade; EXAT = Exatidão.

Parâmetros otimizados modelos RF

A

Conjunto dados	<i>nTrees</i>	<i>mTry</i>	mínimo nó folha
01	1000	40%	2
02	1000	30%	2
03	500	50%	4
05	1500	30%	2
06	1500	30%	2

C

Conjunto dados	<i>nTrees</i>	<i>mTry</i>	mínimo nó folha
01	1000	30%	2
02	500	10%	2
03	1500	5%	2
05	1500	10%	5
06	1500	10%	6

Figura S04 Parâmetros otimizados para modelos de classificação RF a partir de variáveis selecionadas pelo método URF (A) e variáveis espectrais na região de *fingerprint* (C).

Parâmetros otimizados modelos URF-PCoA-RF

A

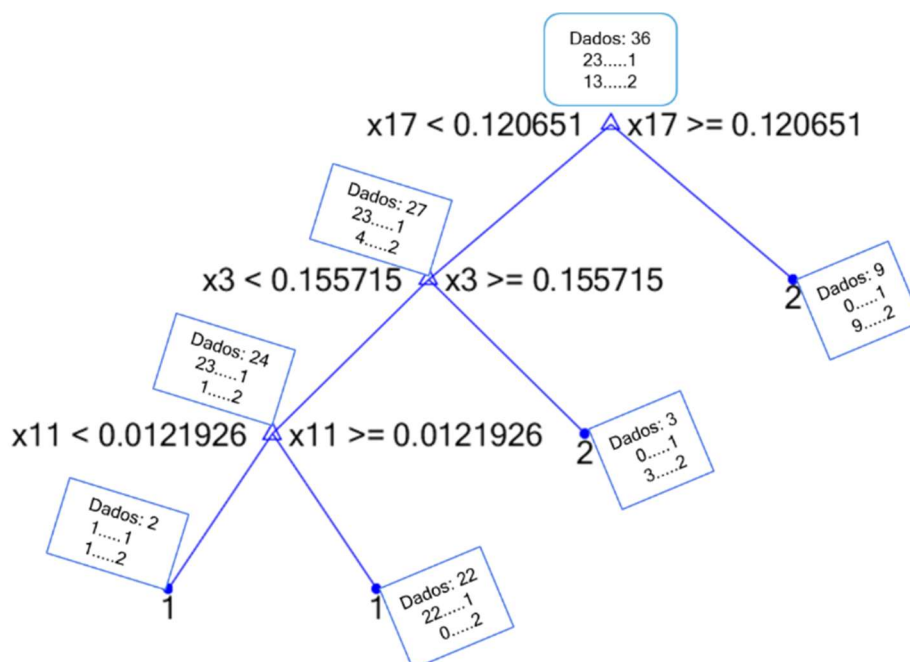
Conjunto dados	<i>nTrees</i>	<i>mTry</i>	mínimo nó folha
01	1000	13%	3
02	1000	14%	3
03	1000	20%	1
05	1000	15%	1
06	1000	17%	1

B

Conjunto dados	<i>nTrees</i>	<i>mTry</i>	mínimo nó folha
01	1000	13%	10
02	1000	14%	2
03	1000	20%	8
05	1000	15%	1
06	1500	10%	2

Figura S05 Parâmetros otimizados para modelos de classificação URF-PCoA-RF a partir de dois métodos de atribuição de classes, para amostras não-rotuladas (A) e rotuladas (B).

7.4. ANEXO A - Exemplo de cálculo de índice de Gini em árvore de classificação



Nesta árvore de um exemplo hipotético, o nó raiz é a variável x_{17} , que representa um número de onda amostrado aleatoriamente dentro do parâmetro m_{try} , já selecionado como nó raiz nesta árvore de classificação. Nesta variável há o teste que é a primeira partição realizada neste exemplo ($x_{17} < 0.120651$ ou $x_{17} \geq 0.120651$), representado pela linha azul na figura abaixo (Figura S01). Esta figura é uma forma ilustrativa de demonstrar as partições dadas na árvore do exemplo.

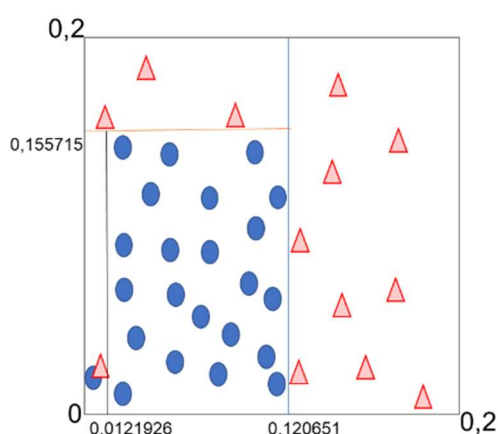


Figura S01 ilustração de distribuição de 36 amostras de exemplo hipotético para árvore de classificação, sendo 23 amostras pertencentes à classe 1 (círculo azul) e 13 amostras pertencentes à classe 2 (triângulo vermelho).

Repare que com a divisão dada pela primeira partição é encontrado um nó folha com 9 amostras que satisfazem o teste $x_{17} \geq 0.120651$, todas pertencentes à

classe 2, não necessitando de mais divisões a partir deste nó. Em outro sentido, as amostras alocadas à esquerda dessa partição ($x_{17} < 0.120651$), precisam de outra divisão, que é dada ortogonal à primeira ($x_3 < 0.155715$ ou $x_3 \geq 0.155715$), representada pela linha laranja na figura S01. Desta nova divisão também resultou um nó terminal (folha) com 3 amostras pertencentes à classe 2 na região que satisfaz o teste $x_3 \geq 0.155715$. Das 36 amostras iniciais, restam então 24, sendo 23 da classe 1 e 1 da classe 2. Uma outra partição dentro do subespaço já particionado ($x_{17} < 0.120651$ e $x_3 < 0.155715$) é realizado com o teste $x_{11} < 0.0121926$ ou $x_{11} \geq 0.0121926$, representada pela linha cinza. Com esta partição é resultado um nó folha contendo 22 amostras pertencentes à classe 1, restando um outro nó folha contendo 1 amostra classe 1 e 1 amostra classe 2.

Para fins didáticos de compreensão do índice de Gini demonstraremos os cálculos com base nesta árvore hipotética. Podemos calcular o índice de Gini para este nó raiz ($x_{17} < 0.120651$ ou $x_{17} \geq 0.120651$) e o ganho de informação dos nós ramos ou descendentes ($x_3 < 0.155715$ ou $x_3 \geq 0.155715$ e $x_{11} < 0.0121926$ ou $x_{11} \geq 0.0121926$), pelos números de registros (amostras) de cada classe em cada nó.

O cálculo do índice de impureza Gini do nó raiz é dado pela Equação 01. Nela são consideradas as frequências relativas de cada classe no respectivo nó.

$$Nó_{raiz(x_{17})} = Classe\ 1 = 23 + Classe\ 2 = 13 \therefore Total = 36$$

$$G_{(nó\ raiz)} = 1 - \left[\left(\frac{23}{36} \right)^2 + \left(\frac{13}{36} \right)^2 \right] = 0,46 \quad \text{Eq. 01}$$

Em seguida, realizamos o cálculo do ganho de informação da variável x_3 , que representa um outro número de onda dentro da amostra aleatória, um nó descendente ao nó raiz. Esse cálculo é realizado em etapas. Primeiro calculamos o índice de Gini para cada teste feito no nó. Se fosse um caso utilizando variáveis categóricas, as respostas no teste seriam Sim ou Não, contabilizando-se os registros para cada resposta do teste. Mas como estamos trabalhando com variáveis numéricas, os registros são encontrados a partir do teste $>$ ou $<$ com base em um valor como parâmetro de divisão recursiva. As equações 02 e 03 calculam então esse índice em cada teste neste nó.

$$Nó_{x_3} = Classe\ 1 = 23 + Classe\ 2 = 4 \therefore Total = 27$$

$$\frac{\sum x_{3 < 0.15}}{\sum \text{nó raiz}} \left(G_{(x_{3 < 0.15})} \right) = \frac{24}{36} \left[1 - \left(\frac{23}{24} \right)^2 - \left(\frac{1}{24} \right)^2 \right] = 0,053 \quad \text{Eq. 02}$$

$$\frac{\sum x_{3 > 0.15}}{\sum \text{nó raiz}} \left(G_{(x_{3 > 0.15})} \right) = \frac{3}{36} \left[1 - \left(\frac{0}{3} \right)^2 - \left(\frac{3}{3} \right)^2 \right] = 0 \quad \text{Eq. 03}$$

A seguir, calculamos o ganho de informação deste nó descendente (Equação 04).

$$G_{(\text{nó raiz})} - G_{(x_3)} = 0,46 - 0,053 - 0 = 0,407 \quad \text{Eq. 04}$$

Da mesma forma, realizamos o cálculo do ganho de informação da variável x_{11} , que representa outro número de onda dentro da amostra aleatória como nó descendente ao nó raiz, com o índice de Gini deste nó nas equações 05 e 06, seguido pelo cálculo de ganho de informação pela Equação 07.

$$Nó_{x_{11}} = \text{Classe 1} = 23 + \text{Classe 2} = 1 \therefore \text{Total} = 24$$

$$\frac{\sum x_{11 < 0.012}}{\sum \text{nó raiz}} \left(G_{(x_{11 < 0.012})} \right) = \frac{2}{36} \left[1 - \left(\frac{1}{2} \right)^2 - \left(\frac{1}{2} \right)^2 \right] = 0,028 \quad \text{Eq. 05}$$

$$\frac{\sum x_{11 > 0.012}}{\sum \text{nó raiz}} \left(G_{(x_{11 > 0.012})} \right) = \frac{22}{36} \left[1 - \left(\frac{22}{22} \right)^2 - \left(\frac{0}{22} \right)^2 \right] = 0 \quad \text{Eq. 06}$$

A seguir, calculamos o ganho de informação deste nó descendente (Equação 04).

$$G_{(\text{nó raiz})} - G_{(x_{11})} = 0,46 - 0,028 - 0 = 0,432 \quad \text{Eq. 07}$$

Dessa forma podemos considerar que a variável x_{11} fornece maior ganho de informação (0,432) do que a variável x_3 (0,407). Podemos observar ainda que os nós com menores índices de Gini são os nós com maior grau de pureza e o nó com índice de Gini zero indica que neste nó todas as amostras foram classificadas como pertencentes a uma só classe e não precisam ser feitas novas divisões deste nó, ou seja, um nó terminal, ou nó folha.

Vale ressaltar que este exemplo utilizou variáveis já selecionadas como preditoras para construção desta árvore diante de uma série de variáveis contínuas. Este exemplo foi demonstrado apenas para melhor compreensão de como o cálculo pode trazer informação da importância da variável.

Todas as etapas apresentadas aqui são realizadas para cada uma das variáveis amostradas em *mtry*, em cada crescimento de árvore, nas quais resultam na seleção das variáveis preditoras para cada árvore no modelo RF.

7.5. Produção Científica

Link de Currículo Lattes: <https://lattes.cnpq.br/2620289110303573>

Artigos publicados

Objeto de identificação digital (DOI): 10.1016/j.lwt.2022.114161

Periódico: Lwt – Food Science and Technology. Vol. 172. 2022, 114161

Número artigo eletrônico: 114161

Título: detection and quantification using ATR-FTIR spectroscopy of whey protein concentrate adulteration with wheat flour

Autores: Matthews S. Martins, Márcia Helena c. Nascimento, Leonardo L. Barbosa, Luciene C. G. Campos, Maneesh N. Singh, Francis L. Martin, Wanderson Romão, Paulo R. Filgueiras, Valério Baraúna.

Objeto de identificação digital (DOI): 10.1016/j.fuel.2021.122527

Periódico: FUEL. ISSN: 0016-2361. Vol. 311. 2022.

Número artigo eletrônico: 122527

Título: Determination of gross calorific value in crude oil by variable selection methods applied to ¹³C NMR spectroscopy

Autores: Elisson Henrique de Paulo; Francine dos Santos; Gabriely S. Folli; Layla Santos Paixão; Márcia Helena Cassago Nascimento; Mariana Moro Kuster; Pedro Henrique da Cunha; Eustáquio V. R. de Castro; Álvaro Cunha Neto; Paulo Roberto Filgueiras.

Objeto de identificação digital (DOI): 10.1016/j.microc.2022.107966

Periódico: Microchemical Journal. ISSN: 0026-265X vol.182. 2022,

Número artigo eletrônico: 107966

Título: Effect of fermentation on the quality of conilon coffee (*Coffea canephora*): Chemical and sensory aspects.

Autores: Bárbara Zani Agnoletti; Willian dos Santos Gomes; Gustavo Falqueto de Oliveira; Pedro Henrique da Cunha; Márcia Helena Cassago Nascimento; Álvaro Cunha Neto; Lucas Louzada Pereira; Eustáquio Vinícius Ribeiro de Castro; Emanuele Catarina da Silva Oliveira; Paulo Roberto Filgueiras.

Objeto de identificação digital (DOI): 10.1021/acs.analchem.1c04162

Periódico: Analytical Chemistry. ISSN: 0016-2361. Vol. 94. 2022, p.2425-2433

Número artigo eletrônico: 122527

Título: Noninvasive Diagnostic for COVID-19 from Saliva Biofluid via FTIR Spectroscopy and Multivariate Analysis

Autores: Márcia Helena Cassago Nascimento; Wena D. Marcarini; Gabriely S. Folli; Walter G. da Silva Filho; Leonardo L. Barbosa; Elisson H. de Paulo; Paula F. Vassallo; José Geraldo Mill; Valério G. Barauna; Francis L. Martin; Eustáquio V. R. Castro; Wanderson Romão; Paulo Roberto Filgueiras.

Objeto de identificação digital (DOI): 10.1111/1556-4029.15011

Periódico: Journal of Forensic Sciences. ISSN: 1556-4029 Vol. 67. 2022, p.1399-1416

Título: Portable Raman spectroscopy Applied to the study of drugs of abuse

Autores: Layla P. Santos, Márcia H. C. Nascimento. Iago H. A. S> Barros, Nayara A. Santos, Valdemar Lacerda Jr., Paulo R. Filgueiras, Wanderson Romão.

Objeto de identificação digital (DOI): 10.1002/cem.3444

Periódico: Journal of Chemometrics. ISSN: 0886-9383. Vol. 36. 2022, e3444

Número artigo eletrônico: 3444

Título: Variable selection by permutation applied in support vector regression models

Autores: Pedro H. P. da Cunha, Ellisson H. de Paulo, Gabriely Silveira Folli, Márcia GH. C. Nascimento, Mariana K. Moro, Paulo R. Filgueiras.

Objeto de identificação digital (DOI) 10.1016/j.vibspec.2021.103299

Periódico: VIBRATIONAL SPECTROSCOPY. ISSN: 0924-2031. Vol. 116. 2021

Número artigo eletrônico: 103299

Título: Use of portable Raman spectroscopy in the quality control of extra virgin olive oil and adulterated compound oils

Autores: Iago H. A. S. Barros; Layla S. Paixão; Márcia Helena Cassago Nascimento; Valdemar Lacerda Jr.; Paulo Roberto Filgueiras; Wanderson Romão.

Objeto de identificação digital (DOI): 10.1002/cem.3231

Periódico: Journal of chemometrics. ISSN: 0886-9383. Vol. e3231. 2020. Pág.1-10

Título: Different strategies for the use of Random Forest in NMR spectra

Autores: Betina Pires Oliveira Lovatti; Márcia Helena Cassago Nascimento; Karla Pereira Rainha; Emanuele C. S. Oliveira; Álvaro Cunha Neto; Eustáquio de Castro; Paulo Roberto Filgueiras.

Objeto de identificação digital (DOI): 10.1016/j.fuel.2020.118462

Periódico: Fuel. ISSN: 0016-2361. Vol. 279. 2020. Número artigo eletrônico: 118462.

Título: Particle swarm optimization and ordered predictors selection applied in NMR to predict crude oil properties

Autores: Ellisson H. de Paulo; Gabriely S. Folli; Márcia H. C. Nascimento; Mariana K. Moro; Pedro H. P. da Cunha; Eustáquio V. R. Castro; Álvaro C. Neto; Paulo Roberto Filgueiras.

Objeto de identificação digital (DOI): 10.1016/j.microc.2020.105544

Periódico: Microchemical Journal. ISSN: 0026-265X vol.159. 2020.

Número artigo eletrônico: 105544

Título: Quantification and classification of vegetable oils in extra virgin olive oil samples using a portable near-infrared spectrometer associated with chemometrics

Autores: Flávia Tosato Borghi; Priscila c. Santos; Francine D. Santos; Márcia H. C. Nascimento; Thayná Correa; Mirelly Cesconetto; André A. Pires; Araceli V. F. N. Ribeiro; Valdemar Lacerda; Wanderson Romão; Paulo R. Filgueiras

Objeto de identificação digital (DOI): 10.1002/cem.3282

Periódico: Journal of Chemometrics. ISSN: 0886-9383. Vol. e3282. 2020.

Número artigo eletrônico: e3282

Título: Variable selection in support vector regression using angular search algorithm and variance inflation factor

Autores: Gabriely S. Folli; Márcia H. C. Nascimento; Ellisson H. de Paulo; Pedro H. P. da Cunha; Wanderson Romão; Paulo R. Filgueiras.

Objeto de identificação digital (DOI): 10.1016/j.microc.2018.12.028

Periódico: Microchemical Journal. ISSN: 0026-265X vol.145. 2019, p.1129-1134.

Título: Use of Random forest in the identification of important variables

Autores: Betina Pires Oliveira Lovatti; Márcia Helena Cassago Nascimento; Álvaro Cunha Neto; Eustáquio V. R. de Castro; Paulo Roberto Filgueiras.

Objeto de identificação digital (DOI): 10.1016/j.fuel.2018.07.050

Periódico: Fuel. ISSN: 0016-2361 vol.234. 2018, p.643-649.

Título: Determination of flash point and Reid vapor pressure in petroleum from HTGC and DHA associated with chemometrics

Autores: Márcia Helena Cassago Nascimento; Betina Pires Oliveira Lovatti; Karla Pereira Rainha; Eustáquio V. R. de Castro; Samantha R. C. Silva; Paulo Roberto Filgueiras.

Capítulo de livro publicado

PAULO, Ellisson H.; NASCIMENTO, M. H. C.; da CUNHA, P. H. PEREIRA, L. L.; OLIVEIRA, E. C. S.; FILGUEIRAS, P. R.; FERRÃO, M. F. Determinação de atributos sensoriais de café por método de seleção de variáveis aplicado a RMN de H. In: As Ciências Exatas e a evolução do conhecimento. 1.ed. Curitiba: reflexão Acadêmica, 2022, v.1, p.86-95.

Resumos expandidos publicados em anais de eventos 2019-2022

- NASCIMENTO, M. H. C.; SANTOS, L. P.; BARROS, I. H. A. S.; PAULO, Ellisson H.; ROMAO, W.; FILGUEIRAS, P. R. Autenticidade de cédulas de segunda família do Real com espectroscopia Raman portátil e modelos one-class In: XXII Encontro Latino Americano de Pós-Graduação (XII EPG), 2022, São José dos Campos. XXII EPG Anais Trabalhos Ciências Exatas e da Terra. São José dos Campos. Universidade do Vale do Paraíba, 2022. v.1. p.1 – 6.
- NASCIMENTO, M. H. C.; Dias, P. A. C.; Silva da, T. B.; Brun, B. F.; BORGES, C. S.; BARAUNA, V. G.; FILGUEIRAS, P. R. Novas abordagens de aprendizagem de máquinas para identificação de variáveis importantes em bioespectroscopia In: XXII Encontro Latino Americano de Pós-Graduação (XII EPG), 2022, São José dos Campos. XXII EPG Anais Trabalhos Ciências Exatas e da Terra. São José dos Campos: Universidade do Vale do Paraíba, 2022. v.1. p.1 – 6.
- MARTINS, MATTHEWS S.; NASCIMENTO, MÁRCIA H. C.; ROMAO, Wanderson; FILGUEIRAS, P. R.; BARAUNA, V. G. Determinação da adulteração de WHey Protein concentrado com farinha de trigo utilizando espectroscopia ATR-FTIR In: XXII Encontro Latino Americano de Iniciação Científica, 2022, São José dos Campos. XXVI Encontro Latino Americano de Iniciação Científica Anais do Evento Ciências da Saúde. São José dos Campos: Universidade Vale do Paraíba, 2022. v.1. p.1 – 6.
- PAULO, Ellisson H.; NASCIMENTO, MÁRCIA H. C.; RECH, A. M.; WEILER, F. H.; FILGUEIRAS, P. R.; Ferrão, Marco Flores. Determinação de proteínas e açúcares em leite de soja por ATR-FTIR e PLS com seleção de variáveis In: XXII Encontro Latino Americano de Pós-Graduação, 2022, São José dos Campos. XXVI Encontro Latino Americano de Pós-Graduação - Anais do Evento Ciências Exatas. São José dos Campos: Universidade Vale do Paraíba, 2022. v.1. p.1 – 6.
- TORRES, A. L. S.; MATOS, G. M.; GOMES, K. C. G. M.; NASCIMENTO, MÁRCIA H. C.; FARIAS, C. C.; FILGUEIRAS, P. R.; BARAUNA, V. G. Estudo do perfil espectral de compostos anticoagulantes e sua influência na bioespectroscopia In: XXII Encontro Latino Americano de Iniciação Científica, 2022, São José dos Campos. XXVI Encontro Latino Americano de Iniciação Científica Anais do Evento Ciências da Saúde. São José dos Campos: Universidade Vale do Paraíba, 2022. v.1. p.1 – 6.
- PAULO, Ellisson H.; NASCIMENTO, MÁRCIA H. C.; RECH, A. M.; WEILER, F. H.; FILGUEIRAS, P. R.; Ferrão, Marco Flores. Identificação de adulteração em leite de soja por adição de água utilizando FTIR e modelo oneclass In: XXII Encontro Latino Americano

de Pós-Graduação, 2022, São José dos Campos. XXVI Encontro Latino Americano de Pós-Graduação - Anais do Evento Ciências Exatas. São José dos Campos: Universidade Vale do Paraíba, 2022. v.1. p.1 – 6.

- PAULO, Elisson H.; NASCIMENTO, MÁRCIA H.C.; CUNHA, Pedro H. P.; PEREIRA, L. L.; OLIVEIRA, E. C. S.; FILGUEIRAS, P. R. Determinação de atributos sensoriais de café por método de seleção de variáveis em RMN de ^1H In: XXI Encontro Latino Americano de Pós-Graduação (XXI EPG), 2021, São José dos Campos. INIC_2021 Anais Trabalhos Ciências Exatas e da Terra. ISSN 9786588226032 São José dos Campos-SP: Universidade do Vale do Paraíba, 2021. v.1. p.1 – 6.
- PAULO, E. H.; FOLLI, Gabriely S.; SANTOS, L. P.; NASCIMENTO, MÁRCIA H.C.; CUNHA, Pedro H. P.; CUNHA NETO, A.; FILGUEIRAS, P. R. Seleção de variáveis aplicada a RMN de ^{13}C para a determinação do poder calorífico superior em petróleo In: XXI Encontro Latino Americano de Pós-Graduação (XXI EPG), 2021, São José dos Campos-SP. INIC_2021 Anais Trabalhos Ciências Exatas e da Terra. ISSN 9786588226032 São José dos Campos-SP: Universidade do Vale do Paraíba, 2021. v.1. p.1 – 7.

Resumos publicados em anais de eventos 2019-2022

- PAULO, E. H.; NASCIMENTO, M. H. C.; FILGUEIRAS, P. R. Aplicação de algoritmos inspirados na natureza como seleção de variáveis em infravermelho médio para caracterização de enxofre em petróleo. In: VIII Encontro Capixaba de Química - SBQ/ES. 2021.
- NASCIMENTO, M. H. C.; PAULO, E. H.; SANTOS, F. D.; MARCARINI, W. D.; BARBOSA, L. L.; BARAUNA, V.; ROMÃO, W.; FILGUEIRAS, P. R. Espectroscopia no infravermelho para diagnóstico de COVID-19 a partir de saliva: Uma comparação entre NIR e MIR. Resumos do evento VIII Encontro Capixaba de Química. 2021.
- SANTOS, L. P.; FOLLI, Gabriely S.; NASCIMENTO, MÁRCIA H.C.; SANTOS, F. D.; BARROS, I. H. A. S.; PAULO, Elisson H.; PIRES, A. A.; FILGUEIRAS, P. R.; ROMAO, Wanderson. Classificação multiclasse para verificação de autenticidade do mel utilizando espectroscopia NIR em equipamento portátil In: 17ª Semana Nacional de Ciência e Tecnologia, 2020, Vila Velha. ANAIS DO EVENTO 17ª Semana Nacional de Ciência e Tecnologia - Campus Vila Velha. , 2020. v.1. p.2 – 2
- NASCIMENTO, MÁRCIA H.C.; ROMAO, Wanderson; FILGUEIRAS, P. R. Estudo de caso de determinação de teor de álcool em cachaça como ferramentas para ensino de Química para estudantes da modalidade de jovens e adultos - um relato de experiência In: 17ª Semana Nacional de Ciência e Tecnologia, 2020, Vila Velha. ANAIS DO EVENTO 17ª Semana Nacional de Ciência e Tecnologia - Campus Vila Velha. 17ª Semana Nacional de Ciência e Tecnologia, 2020. v.1. p.7 – 7
- FOLLI, Gabriely S.; SANTOS, L. P.; PAULO, Elisson H.; CUNHA, Pedro H. P.; SANTOS, F. D.; NASCIMENTO, MÁRCIA H.C.; ROMAO, Wanderson; FILGUEIRAS, P. R. Identificação de adulteração em azeite de oliva extra-virgem por micronir associada à SVM mult-class e SVM one-class com adição de outliers artificiais In: XI Workshop de Quimiometria, 2020, Campina Grande. Anais do XI Workshop de Quimiometria. Campina Grande: XI Workshop de Quimiometria, 2020. p.184 – 184
- FOLLI, Gabriely S.; FILGUEIRAS, P. R.; PAULO, Elisson H.; SANTOS, L. P.; NASCIMENTO, MÁRCIA H.C.; TOSATO, F.; PIRES, A. A.; ROMAO, Wanderson. Identificação de adulteração em azeite de oliva extra-virgem por micronir associada à SVM mult-class e SVM one-class com adição de outliers artificiais In: 17ª Semana Nacional de Ciência e Tecnologia, 2020, Vila Velha. ANAIS DO EVENTO 17ª Semana

Nacional de Ciência e Tecnologia - Campus Vila Velha. 17ª Semana Nacional de Ciência e Tecnologia, 2020. v.1. p.8 – 8.

- NASCIMENTO, MÁRCIA H.C.; FOLLI, Gabriely S.; SANTOS, L. P.; SANTOS, F. D.; SAD, C. M. S.; PAULO, Elisson H.; ROMAO, Wanderson; FILGUEIRAS, P. R. Identificação de adulteração em óleo diesel por meio de espectroscopia IV e quimiometria In: 17ª Semana Nacional de Ciência e Tecnologia, 2020, Vila Velha. ANAIS DO EVENTO 17ª Semana Nacional de Ciência e Tecnologia - Campus Vila Velha. 17ª Semana Nacional de Ciência e Tecnologia, 2020. v.1. p.9 – 9.
- SANTOS, L. P.; FOLLI, Gabriely S.; NASCIMENTO, MÁRCIA H.C.; SANTOS, F. D.; PAULO, Elisson H.; ROMAO, Wanderson; FILGUEIRAS, P. R. Método de detecção de droga precursora do golpe “Boa Noite, Cinderela” por espectroscopia Raman em instrumentação portátil e calibração multivariada In: XI Workshop de Quimiometria, 2020, Campina Grande. Anais do XI Workshop de Quimiometria. Campina Grande: XI Workshop de Quimiometria, 2020. p.187 – 187.
- NASCIMENTO, MÁRCIA H.C.; BORGHI, FLAVIA T.; FOLLI, Gabriely S.; SANTOS, L. P.; SANTOS, F. D.; PAULO, Elisson H.; ROMAO, Wanderson; FILGUEIRAS, P. R. Modelos *one-class* e *outlier* artificial na identificação de adulteração em azeite In: XI Workshop de Quimiometria, 2020, Campina Grande. Anais do XI Workshop de Quimiometria. Campina Grande: XI Workshop de Quimiometria, 2020. p.126 – 126.
- NASCIMENTO, MÁRCIA H.C.; FOLLI, Gabriely S.; SANTOS, L. P.; SANTOS, F. D.; SAD, C. M. S.; PAULO, Elisson H.; ROMAO, Wanderson; FILGUEIRAS, P. R. Uso de FT-MIR e Floresta Aleatória na identificação de adulteração de diesel In: XI Workshop de Quimiometria, 2020, Campina Grande. Anais do XI Workshop de Quimiometria. Campina Grande: XI Workshop de Quimiometria, 2020. p.141 – 141.
- SANTOS, L. P.; FOLLI, Gabriely S.; NASCIMENTO, MÁRCIA H.C.; SANTOS, F. D.; PAULO, Elisson H.; ROMAO, Wanderson; FILGUEIRAS, P. R. Verificação de autenticidade do mel utilizando espectroscopia NIR em equipamento portátil e máquina de vetores de suporte In: XI Workshop de Quimiometria, 2020, Campina Grande. Anais do XI Workshop de Quimiometria. Campina Grande: XI Workshop de Quimiometria, 2020. p.201 – 201.
- FOLLI, Gabriely S.; NASCIMENTO, MÁRCIA H.C.; MORO, Mariana K.; PAULO, Elisson H.; CUNHA, Pedro H. P.; ROMAO, Wanderson; FILGUEIRAS, P. R. Algoritmo de busca angular e fator de inflação de variância aplicados na seleção de variáveis para regressão por vetores de suporte In: IV Escola de Inverno de Quimiometria, 2019, Porto Alegre-RS. IV Escola de inverno de quimiometria - Livro de resumos. Porto Alegre: UFRGS, 2019. v.1. p.40 – 40.
- PAULO, Elisson H.; SANTOS, F. D.; FOLLI, G. S.; MORO, M. K.; NASCIMENTO, MÁRCIA H.C.; CUNHA, Pedro H. P.; CUNHA NETO, A.; FILGUEIRAS, P. R. Análise multivariada de dados aplicada a cromatografia gasosa de alta temperatura para estimativa de propriedades de petróleo In: XVI Semana de Ciência e Tecnologia & IV Jornada Integrada de Educação em Ciências, 2019, Vila Velha. Anais do XVI Semana de Ciência e Tecnologia & IV Jornada Integrada de Educação em Ciências - IFES Vila Velha, 2019.
- NASCIMENTO, MÁRCIA H.C.; SANTOS, L. P.; TOSATO, F.; PAULO, Elisson H.; SANTOS, F. D.; FOLLI, G. S.; CUNHA, Pedro H. P.; ROMAO, Wanderson; FILGUEIRAS, P. R. Análise multivariada de dados de espectroscopia Raman portátil In: XVI Semana de Ciência e Tecnologia & IV Jornada Integrada de Educação em Ciências, 2019, Vila Velha. Anais da XVI Semana de Ciência e Tecnologia & IV Jornada Integrada de Educação em Ciências, IFES Vila Velha. , 2019.
- NASCIMENTO, MÁRCIA H.C.; PAULO, Elisson H.; FOLLI, G. S.; CUNHA, Pedro H. P.;

- OLIVEIRA, B. P.; FILGUEIRAS, P. R. Aplicação de DHA e SIMDIS para obtenção de curva PEV de petróleo In: XVI Semana de Ciência e Tecnologia & IV Jornada Integrada de Educação em Ciências, 2019, Vila Velha. Anais do XVI Semana de Ciência e Tecnologia & IV Jornada Integrada de Educação em Ciências - IFES Vila Velha. , 2019.
- PAULO, Elisson H.; FOLLI, Gabriely S.; SANTOS, L. P.; NASCIMENTO, MÁRCIA H.C.; CUNHA, Pedro H. P.; NETO, ÁLVARO C.; FILGUEIRAS, P. R. Aplicação de RMN e regressão linear múltipla na estimativa de propriedades físico-químicas de petróleo In: VII Encontro Capixaba de Química SBQ/ES, 2019, Vitória. Anais do VII Encontro Capixaba de Química SBQ/ES. Vitória: UFES, 2019. v.1. p.45 – 45.
 - FOLLI, G. S.; PAULO, Elisson H.; NASCIMENTO, MÁRCIA H.C.; CUNHA, Pedro H. P.; C. NETO, D. M.; ROMAO, Wanderson; LACERDA JUNIOR, V.; FILGUEIRAS, P. R. Avaliação da correlação entre variáveis espectrais para calibração multivariada In: VII Encontro Capixaba de Química SBQ/ES, 2019, Vitória. Anais do VII Encontro Capixaba de Química SBQ/ES. Vitória: Universidade Federal do Espírito Santo, 2019. v.1. p.70 – 70.
 - PAULO, Elisson H.; FOLLI, G. S.; SANTOS, L. P.; NASCIMENTO, MÁRCIA H.C.; CUNHA, Pedro H. P.; CUNHA NETO, A.; FILGUEIRAS, P. R. Correlação dos compostos SAP em petróleo com aplicação de diferentes técnicas analíticas e regressão por componentes principais In: VII Encontro Capixaba de Química SBQ/ES, 2019, Vitória. Anais do VII Encontro Capixaba de Química SBQ/ES. Vitória: Universidade Federal do Espírito Santo, 2019. v.1. p.85 – 85.
 - FOLLI, Gabriely S.; NASCIMENTO, MÁRCIA H.C.; MORO, Mariana K.; PAULO, Elisson H.; CUNHA, Pedro H. P.; ROMAO, Wanderson; FILGUEIRAS, P. R. Detecção de outliers em modelos de regressão não linear In: Iv Escola de Inverno de Quimiometria, 2019, Porto Alegre. IV Escola de inverno de Quimiometria - Livro de resumos. Porto Alegre: UFRGS, 2019. v.1. p.64 – 64.
 - C. NETO, D. M.; CUNHA, Pedro H. P.; PAULO, Elisson H.; FOLLI, G. S.; NASCIMENTO, MÁRCIA H.C.; SAD, C. M. S.; FILGUEIRAS, P. R.; LACERDA JUNIOR, V. Determinação de propriedades de cerveja utilizando infravermelho próximo e médio associada modelagem PLS In: VII Encontro Capixaba de Química SBQ/ES, 2019, Vitória. Anais do VII Encontro Capixaba de Química SBQ/ES. Vitória: Universidade Federal do Espírito Santo, 2019. v.1. p.58 – 58.
 - SANTOS, L. P.; BARROS, I. H. A. S.; TOSATO, F.; C. NETO, D. M.; FILGUEIRAS, P. R.; NASCIMENTO, MÁRCIA H.C.; ROMAO, Wanderson. Estudo de cervejas comerciais e artesanais por espectroscopia Rama portátil In: XVI Semana de Ciência e Tecnologia & IV Jornada Integrada de Educação em Ciências, 2019, Vila Velha. Anais do XVI Semana de Ciência e Tecnologia & IV Jornada Integrada de Educação em Ciências - IFES Vila Velha, 2019.
 - NASCIMENTO, MÁRCIA H.C.; LOVATTI, BETINA P.O.; FOLLI, Gabriely S.; MORO, Mariana K.; PAULO, Elisson H.; CUNHA, Pedro H. P.; ROMAO, Wanderson; FILGUEIRAS, P. R. Floresta randômica com seleção de árvores em dados de FT-ICR MS para determinar propriedades de petróleo In: Iv Escola de Inverno de Quimiometria, 2019, Porto Alegre. IV Escola de inverno de Quimiometria - Livro de resumos. Porto Alegre: UFRGS, 2019. v.1. p.87 – 87.
 - NASCIMENTO, MÁRCIA H.C.; PIRES, B.; RAINHA, K.; PAULO, Elisson H.; FOLLI, G. S.; SANTOS, L. P.; CUNHA, Pedro H. P.; FILGUEIRAS, P. R. Fusão de dados de técnicas GC-FID para determinação de ponto de fulgor e pressão de vapor Reid em petróleo bruto In: VII Encontro Capixaba de Química SBQ/ES, 2019, Vitória. Anais do VII Encontro Capixaba de Química SBQ/ES. Vitória: Universidade Federal do Espírito Santo, 2019. v.1. p.46 – 46.

- NASCIMENTO, MÁRCIA H.C.; PIRES, B.; RAINHA, K.; PAULO, Ellisson H.; FOLLI, G. S.; SANTOS, L. P.; CUNHA, Pedro H. P.; FILGUEIRAS, P. R. Fusão de dados de técnicas GC-FID para determinação de ponto de fulgor e pressão de vapor Reid em petróleo bruto In: VII Encontro Capixaba de Química SBQ/ES, 2019, Vitória. Anais do VII Encontro Capixaba de Química SBQ/ES. Vitória: Universidade Federal do Espírito Santo, 2019. v.1. p.46 – 46.
- MORO, Mariana K.; CUNHA, Pedro H. P.; NASCIMENTO, MÁRCIA H.C.; PAULO, Ellisson H.; FOLLI, Gabriely S.; FILGUEIRAS, P. R. Otimização da fusão de alto nível para predição de propriedades de petróleo. In: Iv Escola de Inverno de Quimiometria, 2019, Porto Alegre. IV Escola de inverno de Quimiometria - Livro de resumos. Porto Alegre: UFRGS, 2019. v.1. p.101 – 101.
- PAULO, Ellisson H.; NASCIMENTO, MÁRCIA H.C.; MORO, Mariana K.; FOLLI, Gabriely S.; CUNHA, Pedro H. P.; CUNHA NETO, A.; FILGUEIRAS, P. R. Otimização por enxame de partículas aplicada na estimativa de propriedades físico-químicas de petróleo In: Iv Escola de Inverno de Quimiometria, 2019, Porto Alegre. IV Escola de inverno de Quimiometria - Livro de resumos, 2019. v.1. p.104 – 104.
- CUNHA, Pedro H. P.; NASCIMENTO, MÁRCIA H.C.; MORO, Mariana K.; PAULO, Ellisson H.; FOLLI, Gabriely S.; FILGUEIRAS, P. R. SPA aplicado na seleção de variáveis para regressão por vetores de suporte In: Iv Escola de Inverno de Quimiometria, 2019, Porto Alegre. IV Escola de inverno de Quimiometria - Livro de resumos. Porto Alegre: UFRGS, 2019. v.1. p.112 -112.
- PAULO, Ellisson H.; NASCIMENTO, MÁRCIA H.C.; MORO, Mariana K.; FOLLI, Gabriely S.; CUNHA, Pedro H. P.; CUNHA NETO, A.; FILGUEIRAS, P. R. Vetores informativos aplicados na seleção de variáveis para previsão de propriedades de petróleo In: IV Escola de Inverno de Quimiometria, 2019, Porto Alegre. IV Escola de Inverno de Quimiometria - Livro de Resumos. , 2019. v.1. p.118 – 118.

Trabalhos apresentados em eventos 2019-2022

- Evento: XXXVI Reunião Anual da FESBE 2022 –. 05/08/2022 a 09/08/2022. Título: Plasma analysis of patients with Fabry's disease through ATR - FTIR spectroscopy. Autores: Carolina Teles Barretto, Márcia Helena Cassago Nascimento, Tiago Barcelos Silva, Pedro Augusto Costa Dias, Bruna Ferro Brun, Luciene Cristina Gastalho Campos, Wanderson Romão, Paulo Roberto Filgueiras, Valério Garrone Baraúna. Participação como coautora.
- Evento: XXXVI Reunião Anual da FESBE 2022 –. 05/08/2022 a 09/08/2022. Título: Study of Multiple Sclerosis by Infrared Spectroscopy. Autores: Raí dos Santos Santiago, Marcia Helena Nascimento Cassago Nascimento, Leonardo Barbosa Leal, Paulo Roberto Figueiras, Wanderson Romão, Regina Eliza Albano Vanzo, Carla Carvalho Nascimento, Bruno Batitucci Castrillo, Paula Zago Melo Dias, Valerio Garrone Barauna, Lívia Carla Melo Rodrigues. Participação como coautora.
- Evento: Congresso Brasileiro de Neurologia – 21 a 24 de setembro de 2022. Título: Study of Multiple Sclerosis by Fourier-Transform Infrared vibrational spectroscopy technique. Autores: Raí dos Santos Santiago, Marcia Helena Cassago Nascimento, Leonardo Barbosa Leal, Bruno Batitucci Castrillo, Paula Zago Melo Dias, Paulo Roberto Figueiras, Wanderson Romão, Valerio Garrone Barauna, Lívia Carla de Melo Rodrigues. Participação como coautora.
- Evento: XXXVI Reunião Anual da FESBE 2022 –. 05/08/2022 a 09/08/2022. Título: Comparison between diluted versus non-diluted plasma to identify Fabry disease by FTIR.

Autores: Pedro Augusto Costa Dias, Tiago Barcelos da Silva , Bruna Ferro Brun , Leonardo Barbosa Leal , Márcia Helena Cassago Nascimento , Wanderson Romão , Paulo Roberto Filgueiras , Carolina Teles Barretto, Luciene Cristina Gastalho Campos Luiz , Valério Garrone Barauna. Participação como coautora.

- Evento: XXXVI Reunião Anual da FESBE 2022 –. 05/08/2022 a 09/08/2022. Título: Use of FTIR in the study of biochemical profile of obese patients. Autores: Amanda Motta Bortoli, Marcia Helena Cassago Nascimento, Thafanys Sousa Travezani, Leonardo Barbosa Leal, Paulo Roberto Filgueiras, Wanderson Romão, Gustavo Peixoto Soares Miguel, Valerio Garrone Barauna, Fabiano Kenji Haraguchi. Participação como coautora.
- Evento: XXXVI Reunião Anual da FESBE 2022 –. 05/08/2022 a 09/08/2022. Título: The use of infrared spectroscopy to classify patients with altered D-dimer values. Autores: Bruna Ferro Brun, Pedro Augusto Costa Dias, Márcia Helena Cassago Nascimento, Wanderson Romão, Paulo Roberto Filgueiras, Valério Garrone Barauna. Participação como coautora.
- Evento XXII Encontro Latino Americano de Pós-Graduação (XXII EPG) – Universidade do Vale do Paraíba (evento *on line*). 19 a 21 de outubro de 2021. Título: Novas abordagens de aprendizagem de máquinas para identificação de variáveis importantes em bioespectroscopia. Autores: MARCIA HELENA CASSAGO NASCIMENTO; Pedro Augusto Costa Dias; Tiago Barcelos da Silva; Bruna Ferro Brum; Carolina Teles Barreto. Valerio G. Barauna, Paulo Roberto Filgueiras. Tipo de apresentação: apresentação no formato e-poster.
- Evento XXII Encontro Latino Americano de Pós-Graduação (XXII EPG) – Universidade do Vale do Paraíba (evento *on line*). 19 a 21 de outubro de 2021. Título: Autenticidade de cédulas de segunda família do Real com espectroscopia Raman portátil e modelos one-class. Autores: Marcia Helena Cassago Nascimento; Layla Paixão Santos; Iago Henrique A. S. Barros; Ellisson Henrique de Paulo; Wanderson Romão; Paulo Roberto Filgueiras. Tipo de apresentação: apresentação no formato e-poster.
- Evento XXVI Encontro Latino Americano de Iniciação Científica (XXVI INIC)– Universidade do Vale do Paraíba (evento *on line*). 19 a 21 de outubro de 2021. Título: Estudo do Perfil Espectral de Compostos Anticoagulantes e sua Influência na Bioespectroscopia. Autores: Marcia Helena Cassago Nascimento; Kelly Caroline Gonçalves Monteiro Gomes, Anne Louise Silva Torres, Geovana Melo de Matos, Márcia Helena Cassago Nascimento, Carine Coneglian de Farias, Paulo Roberto Figueiras e Valerio Garrone Barauna. Participação como coautora.
- Evento: VIII Encontro Capixaba de Química - SBQ/ES – Universidade Federal do Espírito Santo (evento *on line*). 17 a 17 de dezembro de 2021. Título: Espectroscopia no infravermelho para diagnóstico de COVID-19 a partir de saliva: Uma comparação entre NIR e MIR. Autores: Marcia Helena Cassago Nascimento; Ellisson Henrique de Paulo; Francine Dalapícola dos Santos; Wena Dantas Marcarini, Leonardo Barbosa Leal, Valerio G. Barauna, Wanderson Romão e Paulo Roberto Filgueiras. Tipo de apresentação: apresentação no formato e-poster.
- Evento: XI Workshop de Quimiometria – Universidade Estadual da Paraíba (evento *on line*). 30 de novembro a 03 de dezembro de 2020. Título: Uso de FT-MIR e Floresta Randômica na identificação de adulteração de diesel. Autores: Márcia Helena C. Nascimento; Gabriely S. Folli; Layla P. Santos; Francine D. Santos; Cristina M. Sad; Ellisson H. de Paulo; Wanderson Romão; Paulo Roberto Filgueiras. Tipo de apresentação: Apresentação oral.
- Evento: XI Workshop de Quimiometria – Universidade Estadual da Paraíba (evento *on line*). 30 de novembro a 03 de dezembro de 2020. Título: Modelos oneclass e outlier artificial na identificação de adulteração em azeite. Autores: Márcia Helena C.

Nascimento; Flávia T. Borghi; Gabriely S. Folli; Layla P. Santos; Francine D. Santos; Ellisson H. de Paulo; Wanderson Romão; Paulo Roberto Filgueiras. Tipo de apresentação: Apresentação oral.

- Evento: 17ª Semana Nacional de Ciência e Tecnologia, Instituto Federal do Espírito Santo, Campus de Vila Velha, ES, 23 a 24 de novembro de 2020. Título: Identificação de adulteração em óleo diesel por meio de espectroscopia IV e quimiometria. Autores: Márcia Helena Cassago Nascimento; Gabriely S. Folli; Layla P. Santos; Francine D. Santos; Cristina M. S. Sad; Ellisson H. de Paulo; Wanderson Romão; Paulo Roberto Filgueiras. Tipo de apresentação: Poster.
- Evento: 17ª Semana Nacional de Ciência e Tecnologia, Instituto Federal do Espírito Santo, Campus de Vila Velha, ES, 23 a 24 de novembro de 2020. Título: Estudo de caso de determinação de teor de álcool em cachaça como ferramentas para ensino de Química para estudantes da modalidade de jovens e adultos - um relato de experiência. Autores: Márcia Helena Cassago Nascimento; Wanderson Romão; Paulo Roberto Filgueiras. Tipo de apresentação: Poster.
- Evento: VII Encontro Capixaba de Química SBQ/ES, Universidade Federal do Espírito Santo, 04 a 07 de novembro de 2019. Título: Fusão de dados de técnicas GC-FID para determinação de ponto de fulgor e pressão de vapor Reid em petróleo bruto. Autores: Márcia Helena Cassago Nascimento; Betina Pires O. Lovatti; Karla Pereira Rainha; Ellisson H. de Paulo; Gabriely S. Folli; Layla Paixão Santos; Pedro H. P. da Cunha; Paulo Roberto Filgueiras. Tipo de apresentação: Poster.
- Evento: IV Escola de Inverno de Quimiometria, Universidade Federal do Rio Grande do Sul, 20 a 23 de agosto de 2019. Título: Floresta randômica com seleção de árvores em dados de FT-ICR MS para determinar propriedades de petróleo. Autores: Márcia Helena Cassago Nascimento; Betina Pires O. Lovatti; Gabriely S. Folli; Mariana K. Moro; Ellisson H. de Paulo; Pedro H. P. da Cunha; Wanderson Romão; Paulo Roberto Filgueiras. Tipo de apresentação: Poster.

Registro de programa de computador

- Detecção e quantificação de adulteração de WPC com FT usando FTIR , 2022, Brasil. Instituição de Registro: INPI - Instituto Nacional da Propriedade Industrial, Número do Registro: BR512022003523-5. Data de depósito: 15/10/2022, Data da concessão: 27/12/2022. Instituição(ões) Financiadora(s): FAPES; CAPES; CNPq. Finalidade: Detecção e quantificação de adulteração de Whey Protein com adição de farinha de trigo.

Trabalhos premiados

- Menção Honrosa ao trabalho intitulado "The use of infrared spectroscopy to classify patients with altered D-dimer values" XXXVI Reunião Anual da FeSBE, LVII Congresso Anual da SBFIs e o I Congresso DOHaD Brasil, 2022. Autores: Bruna Ferro Brun, Pedro Augusto Costa Dias, Márcia Helena Cassago Nascimento, Wanderson Romão, Paulo Roberto Filgueiras, Valério Garrone Barauna.
- Prêmio de melhores trabalhos apresentados na IV EIQ: IV Escola de Inverno de Quimiometria - Instituto de Química - Universidade Federal do RS. 2019. SPA aplicado na seleção de variáveis para regressão por vetores de suporte. Autores: Pedro H. P. da Cunha, Ellisson H. de Paulo, Gabriely Silveira Folli, Márcia GH. C. Nascimento, Mariana K. Moro, Paulo R. Filgueiras.

- Melhores trabalhos apresentados no 19º ENQA. 19º Encontro Nacional de Química Analítica e 7º Congresso Iberoamericano de Química Analítica. 2018. Uso de Floresta Randômica na Identificação de Variáveis Importantes. Autores: Betina Pires Oliveira Lovatti; Márcia Helena Cassago Nascimento; Álvaro Cunha Neto; Eustáquio V. R. de Castro; Paulo Roberto Filgueiras.