UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS EXATAS - ICEx
DEPARTAMENTO DE QUÍMICA

Vinicius Pires Gonçalves

MONITORAMENTO DA ESTABILIDADE OXIDATIVA DE BIODIESEL EMPREGANDO ESPECTROSCOPIA VIBRACIONAL ASSOCIADA A FERRAMENTAS QUIMIOMÉTRICAS

Belo Horizonte
2022

UFMG/ICEX/DQ. 1.527
D. 831

Vinicius Pires Gonçalves

MONITORAMENTO DA ESTABILIDADE OXIDATIVA DE BIODIESEL EMPREGANDO ESPECTROSCOPIA VIBRACIONAL ASSOCIADA A FERRAMENTAS QUIMIOMÉTRICAS

Dissertação apresentada ao Departamento de Química do Instituto de Ciências Exatas da Universidade Federal de Minas Gerais como requisito parcial para a obtenção do grau de Mestre em Química.

Orientadora: Prof.ª Dr.ª Mariana Ramos de Almeida
Coorientadora: Prof.ª Dr.ª Vânya Marcia Duarte Pasa

Belo Horizonte
2022

Ficha Catalográfica

G635m
2022
D

Gonçalves, Vinicius Pires.
Monitoramento da estabilidade oxidativa de biodiesel empregando espectroscopia vibracional associada a ferramentas quimiométricas [manuscrito] / Vinicius Pires Gonçalves. 2022.
77 f. : il., gráfs., tabs.

Orientadora: Mariana Ramos de Almeida.
Coorientadora: Vânya Márcia Duarte Pasa.

Dissertação (mestrado) – Universidade Federal de Minas Gerais – Departamento de Química.
Bibliografia: f. 67-71.
Anexo: f. 72-77.

1. Química analítica – Teses. 2. Biocombustíveis – Teses. 3. Biodiesel – Teses. 4. Oxidação – Teses. 5. Raman, Espectroscopia de – Teses. 6. Espectroscopia de infravermelho – Teses. 7. Quimiometria – Teses. 8. Aprendizado do computador – Teses. I. Almeida, Mariana Ramos de, Orientadora. II. Pasa, Vânya Márcia Duarte, Coorientadora. III. Título.

CDU 043

Elaborada por Márcia Cavalcanti Moreira – CRB6-3847.

UNIVERSIDADE FEDERAL DE MINAS GERAIS

"Monitoramento da Estabilidade Oxidativa de Biodiesel Empregando Espectroscopia Vibracional Associada A Ferramentas Quimiométricas"

Vinícius Pires Gonçalves

Dissertação aprovada pela banca examinadora constituída pelos Professores:

Profa. Mariana Ramos de Almeida - Orientadora
UFMG

Profa. Vânya Márcia Duarte Pasa - Coorientadora
UFMG

Prof. Paulo Roberto Filgueiras
UFES

Prof. Bruno Gonçalves Botelho
UFMG

Belo Horizonte, 20 de dezembro de 2022.

Documento assinado eletronicamente por **Mariana Ramos de Almeida**, **Professora do Magistério Superior**, em 20/12/2022, às 11:33, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Bruno Goncalves Botelho**, **Professor do Magistério Superior**, em 20/12/2022, às 20:01, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Vanya Marcia Duarte Pasa**, **Professora do Magistério Superior**, em 20/12/2022, às 22:57, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Paulo Roberto Filgueiras**, **Usuário Externo**, em 27/12/2022, às 22:46, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

A autenticidade deste documento pode ser conferida no site https://sei.ufmg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **1977769** e o código CRC **319E7F41**.

AGRADECIMENTOS

Agradeço a professora e orientadora Mariana Ramos Almeida pelos ensinamentos, paciência, dedicação e companheirismo durante todos estes anos.

Ao Programa de Recursos Humanos da Agência Nacional do Petróleo, Gás Natural e Biocombustíveis – PRH-ANP, suportado com recursos provenientes do investimento de empresas petrolíferas qualificadas na Cláusula de P, D&I da Resolução ANP nº 50/2015. Em especial a professora Vânya Márcia e o pesquisador Henrique Oliveira por toda a ajuda durante a realização do trabalho.

Aos meus amigos por terem tornado o caminho até aqui mais tranquilo.

Aos órgãos de fomento por possibilitarem o desenvolvimento deste trabalho

A UFMG e em especial ao departamento de Química pelos tempos difíceis, mas enriquecedores que tive aí.

Continue a nadar...

Continue a nadar...

(Dory, Procurando Nemo)

RESUMO

Combustíveis derivados de petróleo são a principal fonte de energia mundial, entretanto, com a crescente preocupação ambiental global, torna-se necessário o uso de novas fontes de combustíveis, mais baratas e ambientalmente corretas. Dentre estas fontes, destaca-se o biodiesel, um biocombustível que vem sendo misturado ao diesel mineral. Um dos maiores problemas do uso de biodiesel é a formação de sólidos provenientes da sua oxidação, que é cada vez mais comum à medida que o teor de biodiesel aumenta na mistura. Atualmente, existem métodos padronizados de análise de combustíveis, tais métodos são, no entanto, dispendiosos. Assim, urge o desenvolvimento de novas metodologias de análise que sejam mais simples, robustas, portáteis, rápidas e com menor custo, permitindo um monitoramento da cadeia produtiva. Neste contexto, o objetivo do trabalho foi desenvolver modelos de classificação para caracterizar o biodiesel puro em duas categorias: conforme e não conforme, em relação à estabilidade oxidativa**,** empregando técnicas espectroscópica**s** (infravermelho e Raman). As amostras foram analisadas por espectroscopia no infravermelho médio com refletância total atenuada (ATR-FTIR), espectroscopia no infravermelho próximo (NIR) com um equipamento portátil e espectroscopia Raman. Dois tipos de modelos foram construídos e avaliados, um modelo linear empregando a análise discriminante linear (LDA) e um modelo não linear, usando o método Floresta aleatória. Para corrigir o desbalanceamento de classes foi utilizada uma estratégia de reamostragem empregando o método Adasyn. O desempenho dos modelos foi avaliado pela matriz de confusão e os parâmetros eficiência e coeficiente de correlação de Matthews foram calculados. Os modelos construídos foram capazes de classificar o biodiesel em conforme e não conforme, sendo o coeficiente de correlação de Matthews superior a 0,8 para todos os modelos. A técnica ATR-FTIR foi a mais promissora para modelos lineares e a espectroscopia Raman para métodos não lineares, ambos com coeficiente de correlação de Matthews de 0,97. Por fim, a utilização de técnicas espectroscópicas associadas à métodos de aprendizados de máquina para a classificação de biodiesel, segundo as normas da ANP, para a estabilidade oxidativa é promissora, permitindo a inspeção de modo direto das amostras de biodiesel B100, sem nenhum preparo, de forma rápida e com detecção *in loco* com o uso de equipamentos portáteis, disponíveis comercialmente para as três técnicas empregadas neste trabalho.

Palavras-chave: Biocombustível. Quimiometria. Random Forest. LDA. Reamostragem.

ABSTRACT

Petroleum-derived fuels are the primary source of energy worldwide, nevertheless, the growing environmental awareness makes it necessary to establish new renewable sources of fuel, cheaper to produce and environmentally friendly. Biodiesel gains emphasis among these sources, as it has been blended with mineral diesel. One of the biggest problems in using of biodiesel is the formation of solids from its oxidation, which is increasingly common as the biodiesel content increases in the blend. Currently, there are standardized methods for fuel analysis, however, such methods are expensive. Thus, it is urgently necessary to develop new analysis methodologies that are simpler, more robust, portable, faster, and less costly, allowing for monitoring of the production chain. In this context, the objective of the work was to develop classification models to characterize pure biodiesel in two categories: compliant and non-compliant, regarding oxidative stability, employing spectroscopic techniques (infrared and Raman). Samples were analyzed by attenuated total reflectance mid-infrared spectroscopy (ATR-FTIR), near-infrared spectroscopy (NIR) with portable equipment, and Raman spectroscopy. Two types of models were built and evaluated, a linear model employing linear discriminant analysis (LDA) and a non-linear model using the Random Forest method. In addition, a resampling strategy employing the Adasyn method was used to correct for class unbalance. The performance of the models was evaluated by the confusion matrix and the parameters efficiency and Matthews correlation coefficient were calculated. The models here constructed were able to classify biodiesel into compliant and non-compliant, and the Matthews correlation coefficient was greater than 0.8 for all models. The ATR-FTIR technique was the most promising for linear models and Raman spectroscopy for non-linear methods, both with Matthews correlation coefficient of 0.97. Finally, the use of spectroscopic techniques associated with machine learning methods for the classification of biodiesel, according to ANP standards, for oxidative stability is promising, allowing the inspection of B100 biodiesel samples directly, without any preparation, quickly and with on-site detection using portable equipment, commercially available for the three techniques employed in this work.

Keywords: Biofuel. Chemometrics. Random Forest. LDA. Resampling.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS

Adasyn	*Adaptive Synthetic Sampling Approach for Imbalanced Learning*
ANP	Agência Nacional de Petróleo, Gás Natural e Biocombustíveis
ASTM	*American Society for Testing and Materials*
ATR	Reflectância total atenuada (*Attenuated Total Reflection*)
FN	Falso negativo
FP	Falso positivo
FTIR	Infravermelho com transformada de Fourier (*Fourier Transform Infrared*)
iPLS	*Interval Partial Least-Squares*
IR	Infravermelho (*Infrared*)
KNN	K-ésimo vizinho mais próximo (K^{th} *Nearest Neighbour*)
LDA	Análise Discriminante Linear *(Linear Discriminant Analysis*)
LEC	Laboratório de Ensaios de Combustíveis
MCC	Coeficiente de correlação de Matthews (*Matthews Correlation Coefficient)*
MIR	Infravermelho médio (*Mid-Infrared*)
MLR	Regressão linear múltipla (*Multiple-linear regression*)
MS	Espectrometria de massas (*Mass Spectrometry*)
MSC	Correção do espalhamento multiplicativo (*Multiplicative Scatter Correction*)
NIR	Infravermelho próximo (*Near-Infrared*)
PARAFAC	*Parallel factor analysis*
PC	Componente principal (*Principal Component*)
PCA	Análise de componentes principais (*Principal Component Analysis*)
PCR	*Principal component regression*
PLS	Mínimos quadrados parciais (*Partial Least Squares*)
PLS-DA	Análise discriminante por mínimos quadrados parciais (*Partial Least Square Discriminant Analysis*)
R^2	Coeficiente de Determinação
RMSEC	Raiz quadrada do erro quadrático médio (*Root Mean Square Error of Calibration*)
RMSEP	*Root mean square error of prediction*
ROC	*Receiver Operator Curve*
RPD	*Ratio performance deviation*

SMOTE	*Synthetic Minority Over-sampling Technique*
SNV	Variação normal padrão (*Standard Normal Variate*)
SVR	Regressão de vetores de suporte (*Support vector regression*)
VLs	Variáveis latentes
VN	Verdadeiro negativo
VP	Verdadeiro positivo

SUMÁRIO

1. INTRODUÇÃO

Desde o século passado combustíveis derivados de petróleo são a principal fonte de energia mundial. Entretanto, o aumento da demanda energética associada a pesquisas que indicam o esgotamento das reservas de petróleo acarretam um aumento nos preços e, com a crescente preocupação ambiental global, são necessárias novas fontes de combustíveis, mais baratas e ambientalmente corretas (RAMOS *et al.*, 2003).

Dentre estas fontes destaca-se o biodiesel, composto por uma mistura entre ácidos graxos de cadeia longa e metil/etil ésteres. O biodiesel pode ser obtido, dentre outras formas, pela reação de triglicerídeos com álcoois (etanol/metanol) na presença de um catalizador ácido ou básico em uma reação de transesterificação. Deste modo o biodiesel pode ser obtido de gorduras animais, óleos vegetais, e até mesmo a partir de óleo de fritura usado (LEUNG; WU; LEUNG, 2010)

Como o biodiesel e diesel convencional apresentam características semelhantes ambos são misturados e utilizados como combustível. Essas misturas são chamadas BX sendo X a fração de biodiesel presente. Para atender demandas globais de redução de gases de efeito estufa tais como a Rio 92, protocolo de Kyoto, Rio+20 entre outras, o Brasil instaurou a resolução CNPE nº16 de 29 de outubro de 2018 que prevê que o diesel nacional seja B15 até 2023 (BRASIL, 2018). No entanto, estudos mostram que estas misturas podem ser realizadas até um certo nível sem modificações ou danos aos equipamentos originalmente planejados para utilizar o diesel convencional (NAIR, 2015).

Um dos maiores problemas do uso de biodiesel é a formação de sólidos em diesel, que é cada vez mais comum à medida que o teor de biodiesel aumenta na mistura BX, com prejuízos financeiros para os usuários (KUMAR, 2017). Diante desse problema, um monitoramento da qualidade do combustível de forma rápida e em toda a cadeia de distribuição permitirá a detecção dos diversos problemas passíveis de existirem e, principalmente, evitará o uso de biodiesel de qualidade inadequada.

Atualmente, as metodologias empregadas para avaliar a qualidade do biodiesel brasileiro estão bem consolidadas, cujas análises devem atender as especificações requeridas pela principal agência reguladora, a ANP (Agência Nacional do Petróleo), através da resolução Nº 45 de 2014 (BRASIL, 2014). Ressalta-se que a ANP é o órgão federal responsável pela regulação das indústrias de petróleo e gás natural e de biocombustíveis no Brasil, que executa

a política nacional para o setor, com foco na garantia do abastecimento de combustíveis e na defesa dos interesses dos consumidores

No entanto, os métodos padronizados são dispendiosos, e a busca por métodos mais rápidos, que permitam análise *in situ* é de grande interesse na área de controle de qualidade. Dessa forma, novas metodologias de análise que sejam mais simples, robustas, portáteis, rápidas e com menor custo são desejáveis permitindo um monitoramento da cadeia produtiva.

Dentre as técnicas analíticas de simples operação e baixa complexidade, tem-se a espectroscopia vibracional (Infravermelho e Raman). A utilização dos espectros vibracionais associados às ferramentas quimiométricas permite ao analista identificar grupos funcionais das moléculas, no caso em questão, o biodiesel e seus contaminantes, assim como a construção de modelos de classificação e regressão. Atualmente é possível encontrar no mercado equipamentos de uso portátil e de custo relativamente acessível, que poderão ajudar na transposição dos resultados desta dissertação em uma tecnologia de relevância para o país.

Neste trabalho, modelos de classificação multivariada foram construídos e validados a partir dos dados de espectroscopia vibracional (infravermelho próximo e médio, Raman) para classificar amostras de biodiesel (B100) do mercado, em conforme e não-conforme quanto à estabilidade oxidativa. Com a metodologia desenvolvida, as amostras poderão ser inspecionadas de modo direto, sem nenhum preparo, de forma rápida, com detecção *in loco*, evitando seu uso quando reprovadas.

2. OBJETIVOS

Este trabalho teve como o objetivo a classificação de amostras de biodiesel (B100) em conforme e não conforme em relação a estabilidade oxidativa de acordo com a ANP baseado em dados de espectroscopia vibracional e métodos de aprendizado de máquina.

2.1 Objetivos específicos

- Obtenção dos espectros na região do infravermelho médio e próximo, e obtenção dos espectros Raman;

- Caracterização dos espectros obtidos;

- Emprego de ferramentas de reamostragem para balanceamento de classes;

- Construção de modelos de classificação multivariada linear (Análise Discriminante Linear - LDA) e não-linear (Floresta Aleatória) empregando os dados de espectroscopia vibracional para discriminar amostras conforme de amostras não conforme;

- Avaliação do desempenho dos modelos por meio das figuras de mérito: Verdadeiro Positivo, Verdadeiro Negativo, Falso Positivo, Falso Negativo, Eficiência, Coeficiente de correlação de Matthews.

- Caracterização espectroscópicas dos modelos obtidos;

- Comparação dos modelos construídos.

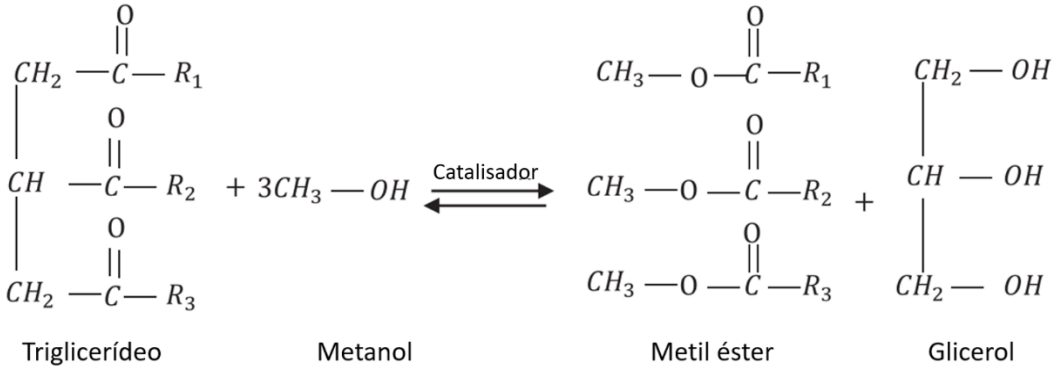
3. REVISÃO DA LITERATURA

3.1 Biodiesel

De acordo com a Resolução 42 de 24/11/2004 da ANP, biodiesel é o combustível formado de mono-alquil ésteres oriundos de gorduras vegetais, animais e óleo de fritura usado, ou seja, é um biocombustível derivado de fontes renováveis (BRASIL, 2004). O biodiesel pode ser obtido por diferentes métodos, tais como reações por fluido supercrítico e transesterificação. O princípio da reação de formação de biodiesel realizada com fluido supercrítico está no efeito da relação entre pressão e temperatura sobre as propriedades do solvente, como viscosidade e densidade específica. Esse método, apesar de ser rápido e não necessitar de catalisador, demanda um alto custo energético e o uso de equipamentos que trabalham em altas pressões e temperaturas, sendo pouco viável economicamente (PELISSON, 2013).

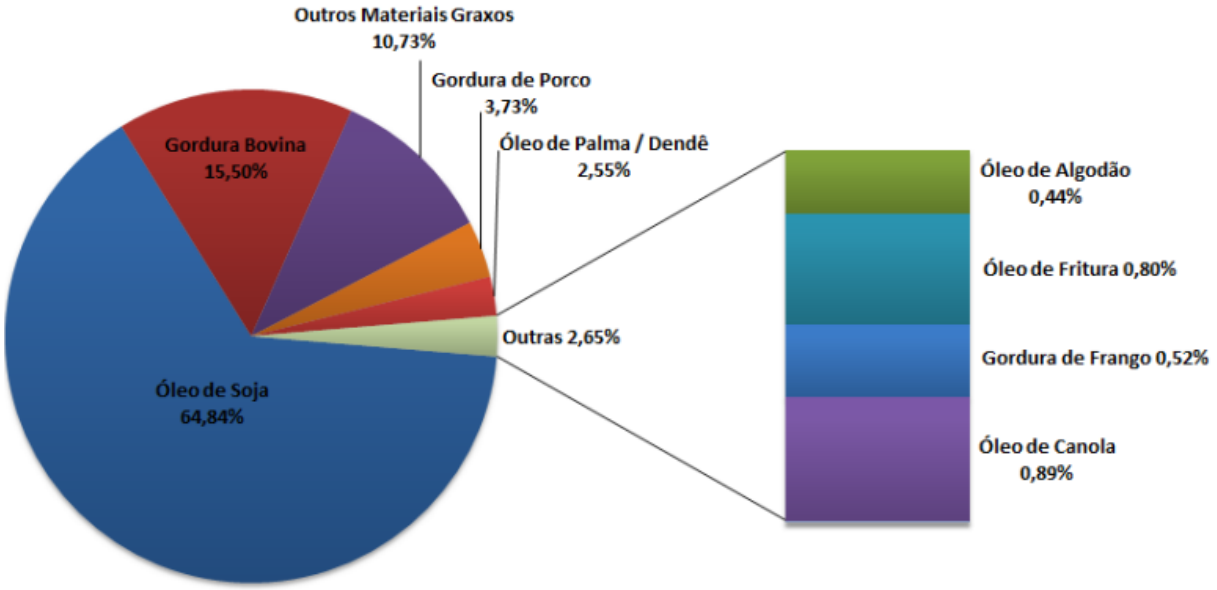
A transesterificação consiste na conversão do óleo animal ou vegetal, com um álcool, podendo ser metanol ou etanol, na presença de catalizador ácido ou básico, sendo o catalisador básico mais utilizado (REZANIA *et al.*, 2019). A reação está descrita pela Figura 1. A transesterificação catalítica é o método mais utilizado, uma vez que seu processo de fabricação é mais barato e mais simples que o método de fluido supercrítico.

Figura 1 – Reação de transesterificação - REZANIA, 2019



Além de métodos alternativos para a obtenção de biodiesel diversas matérias-primas podem ser utilizadas na sua fabricação. Em fevereiro de 2017 o boletim mensal do biodiesel, fornecido pela ANP, informou que as matérias-primas mais utilizadas na produção do biodiesel foram o óleo de soja e o sebo bovino, como pode ser visto pela Figura 2. Observa-se também que em torno de 20% são oriundas de outras fontes, o que mostra a diversidade de origem do produto.

Figura 2 – Perfil nacional de matérias-primas consumidas para produção de biodiesel em janeiro de 2017 – ANP, 2017



O biodiesel pode ser classificado em gerações de acordo com a fonte de lipídeos, o biodiesel de primeira geração é fabricado a partir de óleos comestíveis, como soja e colza. O biodiesel de segunda geração é feito a partir de óleos não comestíveis tais como óleo de jatropha, óleo de semente de seringueira entre outros. Por fim, a fabricação do biodiesel de terceira geração é feita a partir de óleo de micro-algas e óleos de descarte, tais como óleos de fritura usados (SINGH *et al.*, 2020).

O biodiesel faz parte da matriz energética do Brasil desde 2005, sendo adicionado ao diesel mineral a partir da lei n° 11.097/2005 (BRASIL, 2005). Inicialmente utilizou-se o B2 (2% de biodiesel e 98% de diesel mineral em volume) com o aumento gradativo ao longo dos anos. Posteriormente com a lei nº 13.263/2016 (BRASIL, 2016) autorizou o uso de B15 desde que obedecidas as condicionantes de aprovação de testes nos motores para esse teor. A Resolução CNPE n° 16/2018 propôs um cronograma de aumento do percentual de biodiesel na mistura com o diesel de 1% ao ano, atingindo 15%, em 2023 (BRASIL, 2018). Atualmente, o teor de biodiesel acrescido no diesel é de 10% v/v, e esse percentual deve ser mantido até março de 2023.

Com a tendência de alta demanda pelo biodiesel, em virtude do aumento do teor das alíquotas do biodiesel na mistura do diesel para cumprimento das resoluções citadas

anteriormente, aumenta-se também a demanda por análises mais rápidas, de baixo custo e eficientes para aferir a qualidade do biocombustível comercializado.

3.1.1 Controle de qualidade do biodiesel – estabilidade oxidativa

Atualmente, as metodologias empregadas para avaliar a qualidade do biodiesel brasileiro estão bem consolidadas, cujas análises devem atender as especificações requeridas pela principal agência reguladora, a ANP (Agência Nacional do Petróleo), através da resolução Nº 45 de 2014 (BRASIL, 2014).

Devido à grande extensão da cadeia produtiva do biodiesel e de problemas intrínsecos relacionados a não-conformidades que surgem ao longo do tempo, é importante o controle de qualidade em cada etapa da cadeia. Um dos principais problemas do biodiesel se deve à falta de estabilidade oxidativa, levando a formação de sólidos e água na sua composição (KARAVALAKIS; STOURNAS; KARONIS, 2010). A conformidade do biodiesel é sensível ao processo como é preservado, que compreende desde o período da produção, o transporte, o armazenamento na distribuidora, o momento da comercialização e o uso no veículo. Portanto, os agentes causadores de não-conformidades podem estar presentes em todas as etapas da extensa cadeia do biodiesel.

O grande desafio para se ter elevados teores de biodiesel no diesel advém, principalmente, das diferenças entre o diesel fóssil e o biodiesel, cujas moléculas do biocombustível são mais polares pela presença de grupos carboxila. Estas moléculas são mais reativas devido ao grande número de insaturações. O fato torna o biodiesel mais susceptível à degradação por processos oxidativos, térmicos e microbianos. Com a oxidação do biodiesel tem-se a formação de compostos sólidos e água, resultando em entupimentos, aumento dos custos de manutenção e diminuindo o tempo de uso efetivo de equipamentos. Por isso, mesmo com as vantagens ambientais e com o menor custo de produção o biodiesel ainda enfrenta resistência em sua utilização (KUMAR, 2017).

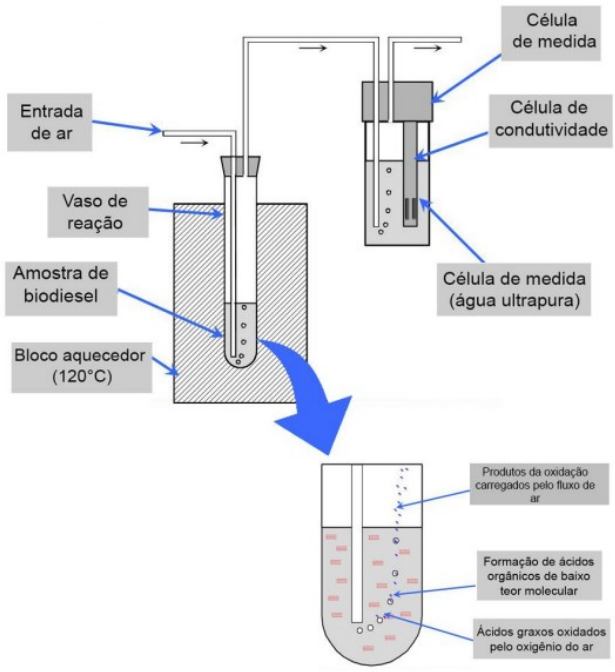
Do Amaral e colaboradores (2020) estudaram a estabilidade oxidativa do biodiesel durante o armazenamento e mostraram que misturas com teores maiores que 10% apresentam baixa estabilidade e, conforme o teor de biodiesel aumenta sua estabilidade decresce. Além disto, os autores apontaram a influência negativa da umidade e temperatura na estabilidade

oxidativa, apresentando um desafio na adoção de valores altos de biodiesel em países tropicais como o Brasil (DO AMARAL; DE REZENDE; PASA, 2020).

McCornick e Westbrook (2010) investigaram a estabilidade oxidativa na armazenagem do biodiesel B100 e de misturas de B5 a B20 por longos períodos utilizando métodos padronizados. Os resultados obtidos indicam que B100 sem a presença de antioxidantes não é estável por mais de três meses, mas com a sua adição a estabilidade pode chegar a seis meses. As misturas B5 e B20 preparadas a partir de B100 em conformidade com a norma EN 14112 apresentam estabilidade de doze e quatro meses respectivamente. Porém, caso o B100 utilizado não esteja dentro dos limites da norma a estabilidade cai bastante podendo ser tão pequena quanto um mês, evidenciando a importância do seu monitoramento (MCCORMICK; WESTBROOK, 2010).

A quantificação da estabilidade oxidativa é feita através de métodos já consolidados baseados em medições de oxidação acelerada. O método do Rancimat, utilizado nas normas EN 14214, ASTM D6751, ANP 45/2014, consiste em aquecer o combustível a 110 ºC em um recipiente selado com o fluxo de ar constante. Produtos voláteis da oxidação, tais como ácido acético e ácido fórmico, são arrastados para uma câmara com água destilada e a condutividade da água é monitorada, a estabilidade oxidativa é determinada pela variação brusca da condutividade, a figura 3 ilustra o esquema do método. Outro método empregado para determinar a estabilidade oxidativa é o Petro-Oxy, que consiste em uma câmara com o biodiesel aquecida a 140 ºC e pressurizada com 700 kPA de O_2 sendo a estabilidade oxidativa correlacionada com o tempo necessário para uma queda de 10% da pressão (MURTA VALLE; LEONARDO; DWECK, 2014). O tempo médio para a obtenção dos resultados via Rancimat é aproximadamente 12h (mínimo aceito como conforme pela ANP é de 12h) mas podendo superar as 20h. Para o Petro-Oxy o tempo de análise é em torno de 3h.

Figura 3 – Esquema do método Racimat – Adaptado de PULLEN, SAEED (2012)



3.2 Espectroscopia Vibracional

3.2.1 Espectroscopia Vibracional na Região do Infravermelho

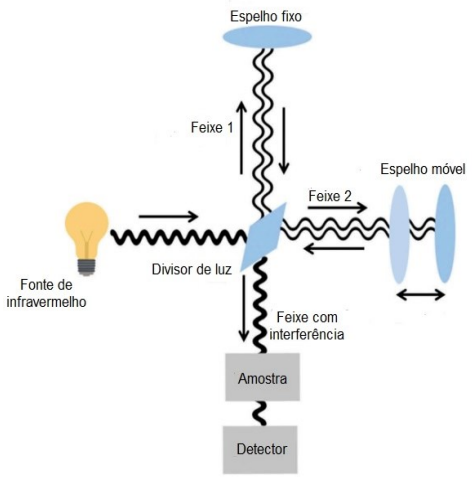
A espectroscopia vibracional no infravermelho consiste em medir a energia absorvida pelas moléculas em estudo, empregando uma fonte de radiação na região do infravermelho do espectro eletromagnético. A região do infravermelho pode ser dividia em três partes, infravermelho distante 400 – 10 cm^{-1}, médio 4.000 – 400 cm^{-1} e próximo 13.000 – 4.000 cm^{-1} (STUART, 2004), nesta seção abordaremos as duas últimas regiões.

A absorção da energia na região do infravermelho é resultado da interação da radiação eletromagnética com a molécula. Entretanto, uma molécula só absorve radiação no infravermelho se ela apresenta uma diferença no seu momento de dipolo durante a vibração, que irá oscilar em ressonância com a radiação. Ou seja, a absorção da radiação só ocorre se a diferença entre dois níveis vibracionais corresponder a mesma energia da radiação incidida. Para a maioria dos compostos orgânicos as transições fundamentais, isto é, aquelas do estado fundamental para o primeiro estado excitado ocorrem na região do infravermelho médio. A

energia absorvida pelas moléculas recebe o nome de absorbância e pode ser relacionada com a quantidade de uma substância que apresente grupo funcional ativo nesta região (VALAND *et al.*, 2020).

Atualmente, os espectrômetros para obtenção de espectros na região do infravermelho são com transformada de Fourier. O funcionamento de um espectrômetro por transformada de Fourier é baseado na obtenção de um interferograma, que depois é convertido para um espectro. O interferômetro mais comum é o de Michelson que consiste em dois espelhos planos no qual um é fixo e outro pode se mover de maneira perpendicular ao plano. Deste modo, a radiação incidente é dividida em dois feixes por um divisor de luz. Um feixe é direcionado no espelho fixo e o outro no espelho móvel, sendo que ambos são refletidos no divisor e interagem. Como o caminho ótico do feixe direcionado ao espelho móvel é diferente daquele direcionado ao espelho fixo as ondas vão interagir hora construtivamente hora negativamente dando origem ao interferograma. A partir do sinal do interferograma é aplicada a transformada de Fourier, uma operação matemática, que converte o sinal para o domínio da frequência, obtendo-se o espectro de infravermelho. A Figura 4 traz um esquema do interferômetro (VALAND *et al.*, 2020).

Figura 4 – Esquema de um aparelho FTIR– Adaptado de VALAND, 2020



As principais vantagens em utilizar um espectrômetro por transformada de Fourier é o aumento da relação sinal-ruído e a velocidade de aquisição dos espectros, proporcionando uma aquisição de um espectro em milissegundos (STUART, 2004).

Em espectroscopia no infravermelho médio é comum se utilizar um acessório de reflectância total atenuada para a análise de substâncias sólidas e líquidas. Este acessório é

baseado no fenômeno de reflectância total e pode ser descrito da seguinte forma: um feixe de radiação infravermelho incide em substância (com índice de refração alto o suficiente para que o fenômeno de reflexão total ocorra) em contato com a amostra. Quando a reflexão total ocorre, na interface entre a substância e a amostra surge uma onda evanescente, que é uma onda estacionária, perpendicular a onda refletida que interage com a amostra sendo absorvida essa onda é chamada de onda atenuada. Um espectro de absorção é obtido assim como na espectroscopia vibracional sem o uso do acessório, entretanto o caminho ótico é dependente do número de reflexões que ocorrem dentro do acessório (HIND; BHARGAVA; MCKINNON, 2001; SCHUTTLEFIELD; GRASSIAN, 2008). O uso de acessório de ATR ampliou as aplicações da espectroscopia no infravermelho e mostra vantagens como facilidade de amostragem, pode ser aplicada para amostras sólidas e líquidas, e muitas vezes as análises podem ser realizadas de forma não destrutiva.

Até aqui foram apresentados os fundamentos da espectroscopia na região do infravermelho médio, na qual os espectros mostram informações das transições vibracionais fundamentais. Considerando a porção mais energética do espectro infravermelho, a região do infravermelho próximo (NIR do inglês *Near Infrared*), não é possível observar as transições fundamentais e sim, sobretons e efeitos de combinação. Como esses efeitos geralmente são associados a transições proibidas os coeficientes de absorção são geralmente 1000 – 10 vezes menores que os coeficientes de absorção do infravermelho médio. Isto faz com que a sensibilidade da técnica seja insuficiente para detectar analitos abaixo de 1 ppm. Além disso, esta é uma técnica pouco seletiva, pois modos vibracionais inativos em energias mais baixas, na região do infravermelho médio por exemplo, podem ser ativos no infravermelho próximo devido as combinações citadas anteriormente. Devido a essas características, amostras complexas como combustíveis, polímeros, medicamentos e alimentos geram espectros complexos com muitas bandas sobrepostas e de difícil interpretação mas, a informação está contida no espectro e com o avanço em técnicas matemáticas e computacionais é possível obter a informação relevante (PASQUINI, 2018).

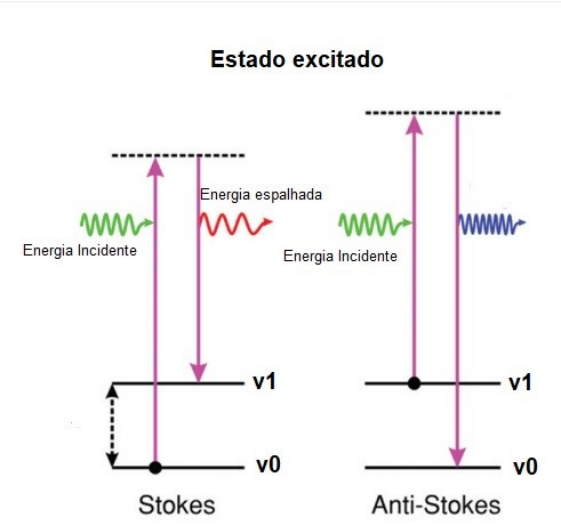
3.2.2 Espectroscopia Raman

Diferentemente da espectroscopia no infravermelho, que mede a radiação absorvida pela molécula, a espectroscopia Raman mede a diferença entre a energia incidida (radiação

monocromática) e a energia espalhada por uma molécula. O espalhamento da radiação pode ser de dois tipos, elástico e inelástico. No primeiro, chamado de espalhamento Rayleigh, não há diferença entre a energia incidida e a espalhada; já no espalhamento inelástico, ocorre a diferença de energia entre a luz incidida e a espalhada, e este tipo de espalhamento caracteriza a espectroscopia Raman.

No espalhamento inelástico a energia espalhada pode ser maior que a energia incidente, neste caso, é chamado de espalhamento anti-Stokes, ou menor, denominada de espalhamento Stokes, conforme esquematizado na Figura 5. Como pode ser observado na Figura 5, a diferença entre a energia incidida e a energia espalhada corresponde a diferença de energia entre dois níveis vibracionais, dessa forma a espectroscopia Raman é um tipo de espectroscopia vibracional, pois fornece informação de modos vibracionais da molécula (HESS, 2021).

Figura 5 – Esquema de espalhamento inelástico - Adaptado de HESS, 2021



Como os fenômenos físicos entre a espectroscopia no infravermelho e Raman são diferentes, a primeira envolve absorção e a segunda envolve espalhamento da radiação, as regras de seleção também são diferentes. Para observar uma banda no infravermelho é necessário que ocorra mudança no momento de dipolo intrínseco da molécula, a absorção é um processo de ressonância, a quantidade de energia incidida deve corresponder a diferença de energia entre dois níveis vibracionais. No caso da espectroscopia Raman, para observar sinais no espectro Raman é necessária uma mudança no momento de dipolo induzido da molécula, conhecido como polarizabilidade (SALA, 2011). Dessa forma, os modos de vibração totalmente simétricos são ativos no Raman e não são no infravermelho, por não possuírem

momento de dipolo intrínseco. As espectroscopias Raman e no infravermelho são conhecidas como técnicas complementares.

Assim como qualquer técnica analítica, a espectroscopia Raman sofreu avanços na sua instrumentação, o que facilitou a tarefa de obter um espectro Raman. Atualmente, equipamentos dispersivos e com transformada de Fourier são comercializados. Os equipamentos dispersivos operam com fonte de excitação na região do UV, visível e início do infravermelho (785 nm), enquanto os equipamentos interferométricos utilizam laser na região do infravermelho próximo (1064 nm) como fonte de excitação (ALMEIDA, 2015).

3.3 Quimiometria

Graças ao avanço tecnológico, espectrômetros e cromatógrafos, equipamentos utilizados em análises de rotina em laboratórios, geram uma quantidade enorme de dados. Um espectro, por exemplo, contém em média 2.000 comprimentos de onda e o resultado de uma única corrida cromatográfica pode gerar cerca de 500.000 dados. Entretanto esse volume de dados não é equivalente a quantidade de informação, logo, é necessário a utilização de novas ferramentas para extrair informações úteis. O emprego de ferramentas matemáticas, estatísticas e computacionais capazes de extrair informações relevantes desse grande conjunto de dados é área da química conhecida como quimiometria (FERREIRA, 2015).

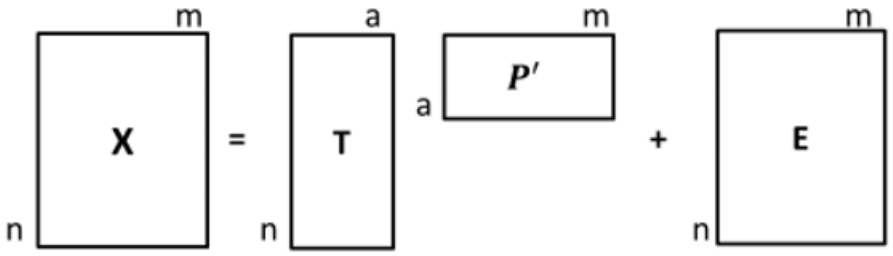
Na quimiometria existem duas grandes áreas, a área de planejamento de experimentos e de modelagem de dados. A área de planejamento de experimentos é voltada para o entendimento e otimização de condições experimentais, ao passo que a área de modelagem de dados busca extrair o máximo de informação dos dados obtidos. Na classificação de modelagem de dados existem duas subdivisões, os métodos supervisionados e os não supervisionados. Os métodos não supervisionados dependem apenas dos dados experimentais e buscam agrupamentos naturais no conjunto de dados. Dentre eles, o mais famoso é a análise de componentes principais (PCA), que é um método de redução de dimensionalidade dos dados de forma que agrupamentos naturais possam ser observados. Neste trabalho foi utilizado a PCA para uma análise exploratória e métodos supervisionados de classificação.

Esta seção apresenta os conceitos fundamentais da análise multivariada com ênfase nos métodos de classificação supervisionados empregados no trabalho.

3.3.1 Análise de componentes principais (PCA)

A PCA tem como objetivo a redução da dimensão dos dados através de combinações lineares de suas variáveis originais. A ideia da PCA é exemplificada na Figura 6.

Figura 6 – Esquema de uma PCA – ALMEIDA, 2015



Os dados instrumentais são representados em **X** com n linhas (amostras) e m colunas (variáveis). Esta matriz é decomposta em uma multiplicação entre as matrizes **T**, contendo os escores, e a matriz **P**, contendo os pesos. A matriz **E** contém os resíduos não explicados pela PCA e deve ser somada à **T** x **P'** para que a igualdade se mantenha.

A matriz **X** é decomposta em uma multiplicação de vetores de dimensionalidade 1, com **t** contendo o escore e **p** contendo os pesos, segundo a equação abaixo.

$$X = t_1 * p_1 + t_2 * p_2 + \mathrm{t_m} * p_a \qquad \textbf{Equação (1)}$$

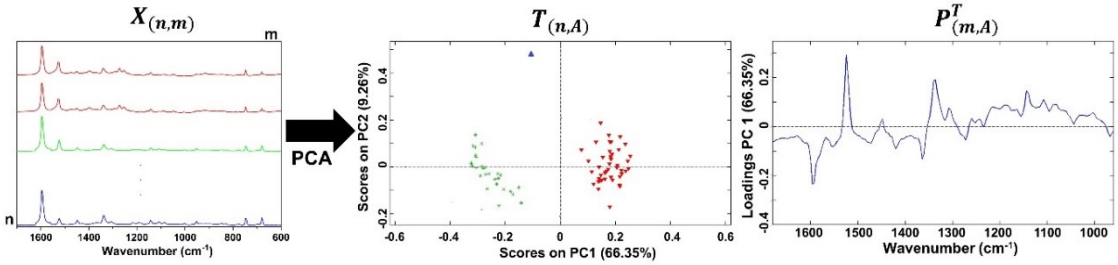
A decomposição de **X** ocorre em a componentes principais (PC), sendo a componente principal uma nova variável gerada a partir da combinação linear de variáveis originais com alta correlação e cada componente principal é construída de modo a explicar a máxima variabilidade dos dados. O vetor peso contém a contribuição de cada variável para as componentes principais e geometricamente representa os cossenos dos ângulos entre a respectiva componente principal e os eixos definidos pelas variáveis originais. O vetor escore contém a projeção da variável original na nova componente principal e o conjunto destes vetores formam as matrizes **T** e **P** (DUNN, 2014).

Na maioria das vezes a PCA é capaz de representar o conjunto de dados em poucas componentes principais, permitindo a visualização de similaridades e diferenças entre as

amostras por meio da análise dos gráficos de escores. As variáveis responsáveis pela similaridade ou diferença entre as amostras podem ser observadas no gráfico dos pesos versus variáveis.

A Figura 7 mostra um exemplo da PCA para dados espectrais (dados contínuos). A matriz **X** (contendo os espectros de n amostras e m variáveis) é decomposta nas matrizes **T** (n, A), contendo os escores, e na matriz **P** (A, m) em a PCs. A relação entre as amostras e os possíveis agrupamentos são observados em um gráfico dos escores das componentes principais um contra o outro.

Figura 7 – PCA para dados espectrais



Na Figura 7 observa-se um grupo de amostras no lado positivo da PC1 (vermelho) e outro grupo no lado negativo da PC1 (verde), indicando que a primeira componente é formada pela combinação de números de onda nos quais estes grupos de amostras apresentam diferenças. Ainda no gráfico de escores observa-se uma amostra com maior valor de escore na PC2 (azul), distante das demais amostras. Neste caso, a PC2 traz informação das características dessa amostra, pois são essas características que a diferem dos dois grupos descritos por PC1. Para o entendimento das diferenças observadas no gráfico de escores é necessário a análise dos pesos (ou *loadings*) das PCs. No gráfico de pesos de uma PC é mostrado quais variáveis contribuem para caracterizar os agrupamentos observados no gráfico de escores. Observe na figura que as bandas espectrais que caracterizam as amostras no lado positivo da PC1 são as bandas próximas a 1500 e 1350 cm^{-1}, enquanto as amostras no lado negativo da PC1, a região espectral com maior peso é em 1600 cm^{-1}.

3.3.2 Aprendizado de Máquina

Recentemente o termo aprendizado de máquina tem ganhado espaço em diferentes áreas do conhecimento. O termo aprendizado de máquina (*machine learning*) é usado para descrever

uma série de métodos que “aprendem com os dados” para construção de modelos que podem tomar decisões com base no que é aprendido (AMIGO, 2021). O termo é frequentemente usado para descrever modelos não lineares ou, ainda de forma mais controvérsia, para descrever modelos não supervisionados. O termo aprendizado de máquina foi empregado neste trabalho como sinônimo de métodos supervisionados, sejam eles métodos lineares ou métodos não lineares de classificação ou regressão.

Os modelos supervisionados dependem de informação além dos dados experimentais, sendo necessário incluir também a variável dependente na construção do modelo. Caso a variável dependente seja discreta, como por exemplo, “conforme” ou “não conforme”, este modelo é denominado modelo de classificação. Caso a variável seja contínua, como por exemplo, 15% de alcaloides, o modelo é denominado modelo de regressão.

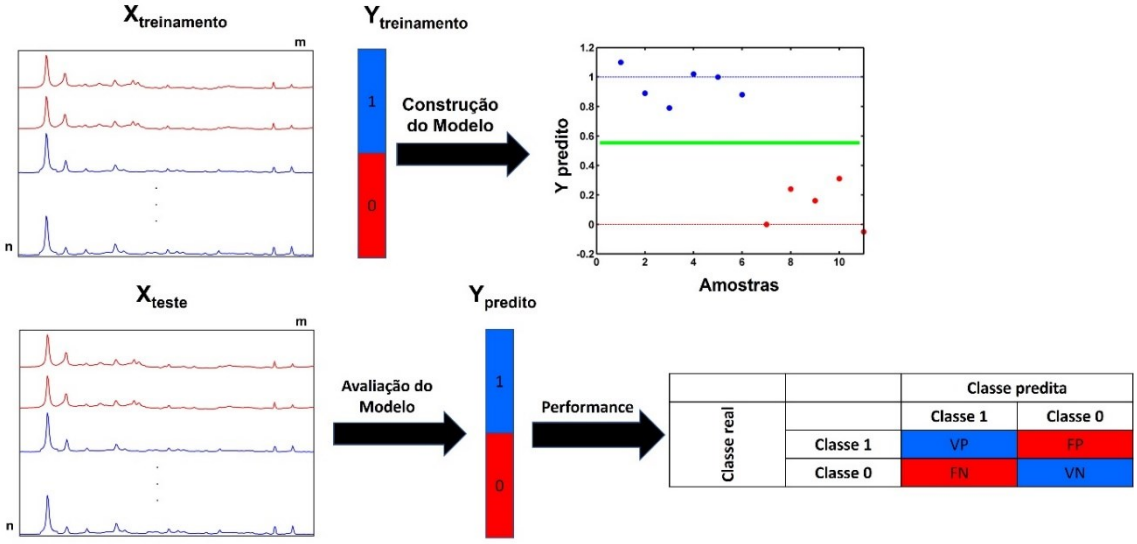
Em problemas supervisionados, são necessários dois conjuntos de dados: um conjunto de treinamento e um conjunto teste. Essa divisão é necessária pois se todas as amostras forem utilizadas para a construção do modelo não conseguimos avaliar sua capacidade em extrapolar a informação aprendida para dados desconhecidos. A falta de capacidade de extrapolação do modelo pode ocorrer de duas maneiras, o modelo pode estar sub ajustado, resultando em um modelo tão simples que é incapaz de explicar a variância dos dados e o modelo pode estar super ajustado, o que significa que ele é tão complexo que explica muito bem os dados aos quais ele foi treinado, mas os resultados para dados desconhecidos são insatisfatórios (MÜLLER, 2018).

A divisão dos dados em conjunto de treinamento e teste é feita de forma que as amostras de treinamento devem representar as variâncias sistemáticas a serem modeladas e distribuídas homogeneamente. Existem vários algoritmos que podem ser utilizados, como Kennard-Stone, Duplex e redes neurais de Kohonen para a seleção das amostras do conjunto de treinamento (WESTAD; MARINI, 2015). Neste trabalho, a amostragem aleatória estratificada foi empregada para divisão do conjunto de dados.

Após a divisão dos dados, os modelos são construídos utilizando o conjunto de treinamento, e validados com o conjunto teste, como citado anteriormente. A validação dos modelos se dá por meio das figuras de mérito que são escolhidas de acordo com a aplicação pretendida para o modelo. Em modelos de classificação geralmente se usa as taxas de verdadeiro positivo, verdadeiro negativo, falso positivo e falso negativo. Além da combinação dessas figuras como eficiência, f1-score, área sobre a curva ROC entre outros (KELLEHER; NAMEE; D’ARCY, 2020).

A Figura 8 ilustra de forma geral o procedimento para a construção de modelos de classificação multivariados. A partir da matriz de dados de treinamento – $\mathbf{X}_{treinamento}$ (por exemplo, dados espectrais) e um vetor **y** (ou matriz **Y**) com a categorização das amostras são calculados os parâmetros do modelo. Com um novo conjunto de dados (matriz de dados $\mathbf{X}_{teste}$), composto por amostras não utilizadas na etapa de treinamento, o modelo é testado (validado) e seu desempenho avaliado através das figuras de mérito.

Figura 8 – Aplicação de um modelo supervisionado



As principais figuras de mérito para os modelos de classificação são baseadas nas taxas de previsões de falso positivo (FP), falso negativo (FN), verdadeiro positivo (VP) e verdadeiro negativo (VN), que podem ser obtidas por meio da matriz de confusão, como mostrado na Figura 9.

Figura 9 – Matriz de confusão

		Classe predita	
Classe real		Classe 1	Classe 0
	Classe 1	VP	FP
	Classe 0	FN	VN

A partir da matriz de confusão, as figuras de mérito taxas de sensibilidade, especificidade e eficiência do modelo são estimadas. Sensibilidade é a taxa de verdadeiros positivos, complementar à taxa FN, enquanto a especificidade é a taxa de verdadeiros negativos,

complementar à taxa FP. Por fim, a taxa de eficiência leva em consideração simultaneamente as taxas FP, FN, VP e VN, como mostrado na equação 2.

$$Eficiência = \frac{VN+VP}{FP+FN+VP+VN} \qquad \textbf{Equação (2)}$$

Uma outra figura de mérito muito utilizada é o coeficiente de correlação de Matthews (MCC), que também é uma combinação das figuras de mérito VP, VN, FP, FN e é calculada pela equação:

$$MCC = \frac{VP \times VN - FP \times FN}{\sqrt{(VP+FP)\times(VP+FN)\times(VN+FP)\times(VN+FN)}} \qquad \textbf{Equação (3)}$$

Que pode ser interpretado da seguinte forma, 1 é uma classificação perfeita, 0 é uma classificação equivalente a considerar todas as amostras pertencentes a uma única classe em problemas binários e -1 é uma classificação com todas as amostras classificadas erradas.

Os métodos supervisionados podem ser lineares e não lineares. Um método é dito linear quando a combinação das variáveis independentes (resposta instrumental) tem uma relação linear com a variável dependente (por exemplo, conforme ou não conforme) e não lineares quando essa relação não é linear. Como representantes de métodos lineares de classificação podemos destacar o LDA (do inglês, *Linear Discriminat Analysis*) e o PLS-DA (do inglês, *Partial Least Squares Discriminant Analysis*). Neste trabalho optou-se pelo uso do LDA que é um dos primeiros métodos classificação, sendo proposto por Sir. Ronald Fischer em 1938 (BLANCHARD; BROWN, 2010). O LDA será discutido com mais detalhes na seção 3.4.3. Como representante dos métodos não lineares optou-se pelo uso do Floresta Aleatória (do inglês, *Random Forest*) proposto por Breiman em 2001 sendo aplicado nas mais diversas áreas do conhecimento com sucesso. Uma discussão mais detalhada sobre o funcionamento do método será dada na seção 3.4.5 (BREIMAN, 2001; LOUPPE, 2014; MÜLLER, 2018).

3.3.3 Análise discriminante linear (LDA)

Existem duas abordagens principais para o método LDA, uma que faz uso do teorema de Bayes e assume uma distribuição de probabilidade normal, e a de Fischer, adotada neste trabalho, que não assume nenhum tipo de distribuição de probabilidades.

Para problemas binários, como os tratados aqui, o método de Fischer pode ser descrito como a projeção dos dados multivariados em uma única função discriminante ($b_{Fischer}$), de modo

que a separação entre o centro das amostras das classes ($\bar{x}$) seja máxima e que a separação entre as amostras de uma mesma classe seja mínima. Para que isso ocorra $b_{Fischer}$ deve ser:

$$b_{Fischer} = S_p^{-1}(\bar{x}_1 - \bar{x}_2) \qquad \textbf{Equação (4)}$$

na qual S_p^{-1} é a matriz agregada de covariância definida como:

$$S_p = \frac{(n_1-1)s_1^2 - (n_2-1)s_2^2}{n_1+n_2-2} \qquad \textbf{Equação (5)}$$

Com n_1 e n_2 sendo o número de amostras nas classes 1, 2 respectivamente e s_1^2 e s_2^2 a variância das classes 1 e 2. Com isso é calculado um escore descriminante y_i para cada amostra da forma:

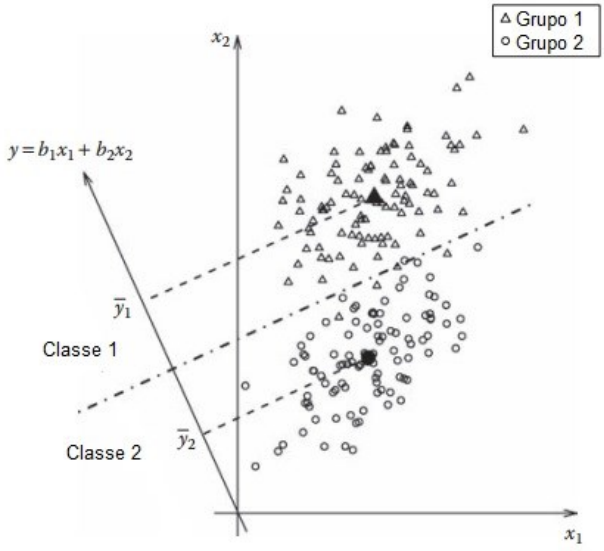
$$y_i = {b_{Fischer}}^T x_i \qquad \textbf{Equação (6)}$$

Este escore é comparado com um limite y_1 calculado da forma:

$$y_1 = \frac{{b_{Fischer}}^T \bar{x}_1 + {b_{Fischer}}^T \bar{x}_2}{2} \qquad \textbf{Equação (7)}$$

Caso $y_i > y_1$ a amostra é classificada como pertencente a classe 2, caso contrário a amostra é classificada como pertencente a classe 1. A Figura 10, a seguir, ilustra o funcionamento do método (BLANCHARD; BROWN, 2010).

Figura 10 – Exemplo de LDA - Adaptado de BROWN, 2010



3.3.4 Árvore de decisão

Árvore de decisão é um método supervisionado que pode ser aplicado tanto a problemas de regressão como a problemas de classificação. Ela é a base do método de Floresta Aleatória (*Random Forest*), e seu entendimento é essencial para a compreensão da Floresta Aleatória.

O objetivo de uma árvore de decisão é subdividir o conjunto amostral em conjuntos menores de forma a minimizar um erro, ou maximizar uma função ganho com comparações do tipo maior, menor, maior ou igual e menor ou igual. Alguns exemplos de erros comumente utilizados são: a raiz quadrada do erro médio, o erro relativo e a impureza Gini. E como exemplos de função ganho podemos citar: a entropia, a área sobre a curva ROC e a eficiência. A escolha de qual erro ou ganho a ser utilizado é selecionado de acordo com o problema em estudo. Aqui o erro escolhido foi a impureza Gini, que é a probabilidade de uma nova amostra ser incorretamente classificada no novo subgrupo. A impureza Gini é matematicamente descrita como o somatório do produto da frequência relativa p de uma classe j no subconjunto l pela frequência relativa das demais classes em cada subconjunto l,

$$Gini = \sum_{j=1}^{k} p_{lj}\,(1 - p_{lj}) \qquad \textbf{Equação (8)}$$

e a frequência relativa (p_{lj}) é dada pela equação abaixo,

$$p_{lj} = \left(\frac{1}{n_l}\right) \sum_{i=1}^{n} I_n \qquad \textbf{Equação (9)}$$

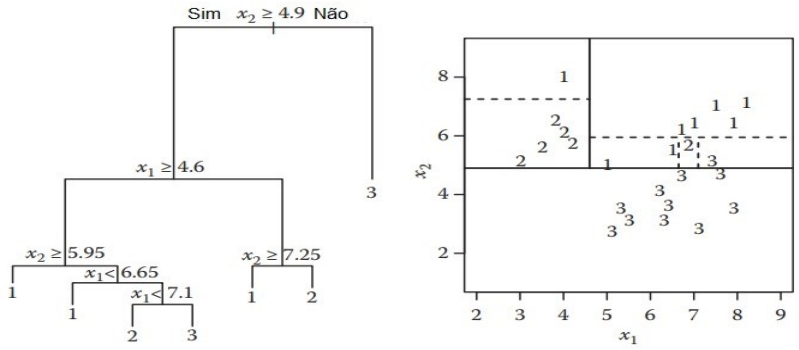
na qual n_l, é o número de amostras no subconjunto l e I é dada por:

$$I \begin{cases} 1, quando\ x_i = j \\ 0, quando\ x_i \neq j \end{cases} \qquad \textbf{Equação (10)}$$

e x_i é uma amostra pertencente ao subconjunto l.

A divisão do conjunto de dados segue de maneira sucessiva de modo semelhante a uma árvore, como ilustrado pela Figura 11, até que um critério de convergência seja satisfeito ou que todos os subconjuntos sejam puros. Cada ponto de divisão é chamado de nó, e os subconjuntos são chamados de galhos, com exceção dos subconjuntos finais que são chamados de folhas.

Figura 11 – Esquema de uma árvore de decisão, a esquerda, originada do conjunto à direita, com classes 1, 2, 3 – Adaptado de BLANCHARD; BROWN, 2010



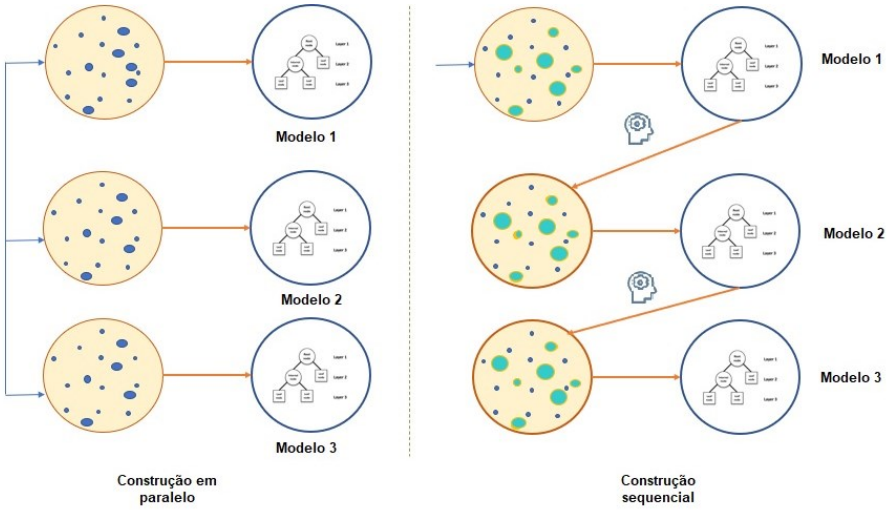
Uma das fragilidades deste método é sua instabilidade, pois uma pequena diferença no valor de uma variável pode levar a árvores completamente distintas. Além disto há a tendência de sobre ajustar o conjunto de treinamento, uma vez que o número de variáveis normalmente é maior que o número de amostras. Deste modo é necessário limitar a complexidade da árvore através de "podas" que podem ser a profundidade da árvore, o número mínimo de amostras em cada subgrupo, o número de variáveis que podem ser consideradas antes de cada subdivisão entre outros (BLANCHARD; BROWN, 2010; MÜLLER, 2018).

3.3.5 Floresta Aleatória

Para contornar essas fragilidades pode-se utilizar uma combinação de árvores de decisão pois, pelas características descritas acima, cada árvore individual tende a ser diferente das outras e a sobre ajustar o conjunto de dados de modo um pouco diferente. Deste modo combinação de árvores distintas compensa os erros das árvores individuais levando a um modelo mais robusto (BLANCHARD; BROWN, 2010).

Existem diversas formas de combinar as árvores para chegar a um modelo mais robusto e/ou eficiente. Duas formas muito comuns, que são a base para a maioria dos métodos, são as baseadas em *baggin* (abreviação para, em inglês, bootstrap aggregation), que consiste em criar modelos independentes em conjuntos de treinamentos artificiais, isto é, em conjuntos Bootstrap, que são conjuntos gerados a partir da reamostragem aleatória com reposição do conjunto original e os modelos de boosting, que consistem em criar modelos sequenciais aumentando o peso das amostras mais difíceis de classificar, conforme ilustrado na Figura 12 (SUTTON, 2005).

Figura 12 – Ilustração Bagging (esquerda) vs Boosting (direita) – Adaptado de SINGHAL, 2022



A floresta aleatória é um método baseado em bagging, no qual a classificação final das amostras é realizada por meio de votação da classificação de cada uma das árvores individuais. Além da reamostragem aleatória do conjunto de dados uma nova camada de aleatoriedade é adicionada ao modelo, pois cada árvore só é exposta a uma porção aleatoriamente escolhida do conjunto, garantindo que as árvores sejam independentes e que com isso a instabilidade dos resultados causada pela variância dos dados seja diminuída (BREIMAN, 2001).

Pela natureza do método a interpretação da importância de cada variável não é direta, uma vez que cada árvore no modelo foi exposta a variáveis diferentes. Por isso uma estratégia para a obtenção da importância das variáveis é o decréscimo médio da impureza Gini em tradução livre (do inglês, *Mean Decrease in Gini impurity*), que é uma média de quanto a impureza Gini diminuí no subconjunto quando aquela variável foi selecionada ponderada pelo número de vezes que a variável foi selecionada (LOUPPE, 2014), deste modo a importância da variável v é matematicamente igual a:

$$Imp(v) = \frac{1}{k}\sum_{\alpha}\sum_{\theta}\Delta_i \qquad \textbf{Equação (11)}$$

na qual $\alpha \in (1,2,\dots,k)$ é associado a cada árvore na floresta, θ são as árvores que contém a variável v em um de seus nós, e Δ_i é a diminuição da impureza Gini ao criar um nó em v.

3.3.6 K-vizinhos mais próximo (KNN)

O KNN, K- vizinhos mais próximos (do inglês, *K Nearest Neighbors*), é um algoritmo que pode ser usado tanto em aprendizado supervisionado, como em aprendizado não supervisionado. Neste trabalho usamos a versão supervisionada de maneira indireta, pois ele é usado em uma etapa intermediária do Adasyn, algoritmo utilizado no trabalho para a correção do desbalanceamento de classes.

Este método é fundamentado na premissa que amostras de um mesmo grupo apresentam um comportamento semelhante em relação a suas variáveis, formando uma vizinhança. A classificação de novas amostras é realizada por meio de votação para problemas de classificação ou a média dos vizinhos para problemas de regressão, sendo que o número de vizinhos mais próximos, k, é determinado experimentalmente. Existem várias métricas para a distância entre vizinhos, as mais comuns são a distância Euclidiana, a distância Manhattan e a distância de Mahalanobis (GUO *et al.*, 2003; MÜLLER, 2018). Neste trabalho utilizamos a distância euclidiana, D, que é a distância utilizada no método do Adasyn e pode ser calculada para dois pontos, y, z, em um espaço n dimensional da seguinte forma:

$$D = \sqrt{\sum(y_i - z_i)^2} \; com \; i \in (1,2,\dots,n) \qquad \textbf{Equação (12)}$$

3.3.7 Método de Reamostragem – Adasyn

Problemas de classificação com classes altamente desbalanceadas apresentam mais uma camada de complexidade quando comparados aos problemas com classes balanceadas. Isso ocorre pois, a maioria dos algoritmos ficam enviesados para classe majoritária (KOVÁCS, 2019).

O problema do desbalanceamento de classes pode ser explicado por meio de um exemplo simples. Considerando um problema binário com 90 amostras pertencentes a classe A e 10 amostras pertencentes a classe B. Ao avaliar o desempenho de um modelo que classifica todas as amostras como pertencentes a classe A, através das figuras de mérito comumente utilizadas, temos: VP = 90; VN = 0; FP = 10; FN = 0; o que resulta em uma eficiência de 90% um resultado que não condiz com o desempenho do modelo. Um problema semelhante ocorre em algoritmos baseados em árvores de decisão nas quais a função custo a ser minimizada (ou o ganho a ser maximizado) pode não ser sensível ao desbalanceamento de classes como é o

caso da eficiência neste exemplo. Além disso, pensando na vizinhança das amostras como no algoritmo do KNN a probabilidade de que a maioria dos vizinhos mais próximos pertencentes a uma amostra aleatória seja pertencente a classe A é maior, simplesmente pela falta de representantes da classe B (BATISTA; PRATI; MONARD, 2004).

Uma abordagem que pode ser utilizada para contornar essa fragilidade é a criação de amostras sintéticas da classe minoritária. Existem diversos métodos para tal, sendo os mais comuns o SMOTE (do inglês *Synthetic Minority Over-sampling Technique*) e suas variações como K-SMOTE, SVM-SMOTE, Borderline-SMOTE e o Adasyn (do inglês *Adaptive Synthetic Sampling Approach for Imbalanced Learning*). Não existe consenso de qual é o método mais adequado, sendo determinado empiricamente (ELREEDY; ATIYA, 2019).

O Adasyn busca resolver o problema de dois modos diferentes, pois aumenta o número de amostras da classe minoritária pela criação de amostras sintéticas. E, além disto, as amostras sintéticas são criadas na região de amostras duvidosas proporcionando mais informação para os modelos na região de dúvida (HE, H., BAI, Y., GARCIA, E., & LI, 2008).

O método pode ser aplicado da seguinte forma, inicialmente define-se um parâmetro β ∈ (0,1], que é o limite para o desbalanceamento aceitável, sendo 1 um conjunto perfeitamente balanceado. Calcula-se o número de amostras G que devem ser criadas artificialmente para atingir o grau de desbalanceamento β da forma:

$$G = \left(m_{majoritárias} - m_{minoritárias}\right) \times \beta \qquad \textbf{Equação (13)}$$

na qual m é o número de amostras pertencentes a cada classe. Para cada amostra X_i pertencente a classe minoritária calcula-se a dificuldade de classificação r_i que é obtida observando-se os k vizinhos mais próximos da forma que,

$$r_i = \Delta/k \qquad \textbf{Equação (14)}$$

$\Delta =$ Número de vizinhos pertencentes a classe **majoritária**

$i \in (1,2,\dots,m_{minoritaria})$

r_i é normalizado para que seja uma função distribuição de densidade $\hat{r}$ da forma,

$$\hat{r}_i = \frac{r_i}{\sum_{i=1}^{m_{minoritarias}} r_i} \qquad \textbf{Equação (15)}$$

Deste modo é possível calcular a quantidade de amostras sintéticas a serem geradas, g_i, para cada amostra X_i pela fórmula,

$$g_i = \hat{r}_i \times G \qquad \textbf{Equação (16)}$$

As amostras sintéticas, s_i, são geradas a partir de uma outra amostra da classe minoritária presente nos K vizinhos escolhida ao acaso, X_{zi}, da forma,

$$s_i = X_i + (X_i - X_{zi}) \times \lambda \qquad \textbf{Equação (17)}$$

sendo $\lambda \in [0,1]$ escolhido aleatoriamente.

3.4 Estudos de biodiesel empregando espectroscopia vibracional e ferramentas quimiométricas

Existe uma busca contínua pelo desenvolvimento de metodologias analíticas para o monitoramento da qualidade de combustíveis. Os métodos analíticos preconizados são os métodos físico-químicos de análise, no entanto, muitos possuem a desvantagem de um longo período de análise. Nas últimas duas décadas métodos espectroscópicos têm sido empregados para a análise de combustíveis, com destaque para os biocombustíveis, juntamente com ferramentas de análise multivariada. Uma consulta na base de dados *Web of Science* pelas palavras "Biodiesel", "*Quality*" e "*Spectroscopy*" no período de 2002 a 2022 retorna 452 artigos. O Brasil é o líder em publicações com 156 artigos, a Índia ocupa a segunda colocação com 52 artigos.

Nesta seção é apresentado uma revisão dos trabalhos encontrados na literatura para o monitoramento da estabilidade oxidativa de biodiesel empregando espectroscopia vibracional, em conjunto com ferramentas quimiométrica.

De Lira e colaboradores (2010) realizaram estudos empregando espectroscopia na região do infravermelho médio e próximo com o uso de ferramentas quimiométricas para a determinação de três parâmetros de qualidade em biodiesel (B100), estabilidade oxidativa, acidez total e teor de água. Três métodos quimiométricos foram utilizados o PLS, iPLS e MLR. O biodiesel foi preparado em laboratório e envelhecidos segundo a ASTM D 4625 por seis semanas. Os combustíveis desse estudo têm como óleos precursores a soja, rabanete, sebo bovino com e sem a adição de antioxidantes. Para a estabilidade oxidativa os modelos baseados em espectroscopia no infravermelho próximo ($R^2 > 0,94$) apresentaram resultados superiores aos modelos de espectroscopia no infravermelho médio ($R^2 > 0,86$). O mesmo comportamento foi observado nos modelos de acidez total, nos quais os modelos baseados em NIR

apresentaram $R^2 > 0,94$ e modelos baseados em MIR $R^2 > 0,92$. Para os modelos de teor de água foram obtidos desempenhos equivalentes, $R^2 = 0,99$. (DE LIRA *et al.*, 2010).

Um modelo para o tratamento de dados de dimensões superiores (PARAFAC), foi utilizado para a identificação das regiões do espectro infravermelho mais relevantes para os testes padronizados de estabilidade. Foram coletadas dezoito amostras de B7 de diversas regiões do Brasil. As amostras foram envelhecidas segundo as metodologias ASTM D4625, ASTM D7545, D6468, CEN EN 15751. Foram obtidos espectros FTIR antes e depois da oxidação das amostras e, através do método PARAFAC, foi possível identificar as regiões dos espectros que mais variaram nos processos de oxidação, exceto nas amostras submetidas a metodologia D6468 (SKROBOT; DE SOUSA SANTOS; BATISTA BRAGA, 2019).

Correia e colaboradores (2018) utilizaram dados de espectroscopia no infravermelho próximo obtidos por um equipamento portátil utilizando o método de regressão PLS para a determinação do teor de biodiesel em diesel, quantidade de enxofre em diesel, quantidade de gasolina, etanol e metanol em gasolina tipo C e teor de água, metanol, etanol em etanol hidratado. Foram preparadas 181 amostras de combustíveis, obtendo excelentes resultados. O modelo de detecção de biodiesel em diesel obteve um RMSEP 1,8% m/m; para o modelo de detecção de enxofre o RMSEP foi de 13,2 mg/L; para os modelos aplicados a gasolina de tipo C os valores de RMSEPs obtidos foram: 0,81%m/m para o teor de gasolina, 3,81% m/m para o teor de etanol e 1,80% m/m para o teor de metanol (CORREIA *et al.*, 2018).

Velvarská e colaboradores (2019) também empregaram a espectroscopia no infravermelho próximo associada a quimiometria para determinar a estabilidade oxidativa de biodiesel e suas misturas. Foram avaliados biodieseis oriundos de colza, girassol e óleo de fritura usado nas proporções de 100%, 30%, 20% e 7%. O método de referência utilizado para a determinação da estabilidade oxidativa foi o Petro-Oxy. Foi utilizado um equipamento com sonda de CaF_2, com um detector de InGaAs, resolução de 8 cm^{-1} e faixa de trabalho de 10,000 – 4,000 cm^{-1}. O método quimiométrico escolhido foi o PLS, com 7 variáveis latentes. O grupo determinou a região ótima para o modelo entre 10,000 – 6,450cm^{-1} e o modelo foi avaliado pelas figuras de mérito RMSEC, R^2, RPD obtendo bons resultados, RMSEC de 8,73 min, R^2 de 0,9600, RPD 3,57, sendo adequado para a determinação da estabilidade oxidativa do biodiesel (VELVARSKÁ *et al.*, 2019).

Recentemente, o SVR (Regressão de vetores de suporte), utilizado para construção de modelo não linear, foi aplicado em dados de FTIR para prever a viscosidade cinemática, o número de cetano, o poder calorífico e a porcentagem de biodiesel no diesel convencional.

Além disto modelos lineares, PLS e PCR também foram testadas para a solução dos mesmos problemas. Os modelos foram construídos com biodiesel de diferentes origens dentre elas: óleo de coco, óleo de palma, óleo de jatropha e óleo de fritura usado. As amostras continham um teor de biodiesel que variava 0 – 100% com incremento de 5%. Para a determinação de biodiesel em diesel os modelos obtiveram excelentes resultados com RMSEP de 0,8%, 2,95% e 3,75% para os modelos SVR, PLS e PCR respectivamente. Para a viscosidade cinemática os RMSEPs obtidos foram: 0,1 ± 0,3%; 0,1 ± 0,4% e 0,8 ± 1,2% para os modelos SVR, PLS, PCR. Essa mesma tendência é observada para os demais parâmetros estudados. Desta forma neste o estudo o SVR se mostrou mais adequado que os modelos construídos com os métodos lineares (BUKKARAPU; KRISHNASAMY, 2022).

Tendo em vista os bons resultados obtidos na literatura para a predição de parâmetros físico-químicos de biodiesel com dados espectroscópicos e quimiometria. Este trabalho teve como o objetivo a avaliação da aplicação de técnicas espectroscópicas (NIR, ATR-FTIR e Raman) e quimiometria em amostras advindas do mercado, sem o preparo de amostras de em laboratório. A maior parte das amostras do mercado possuem conformidade em relação a estabilidade oxidativa, para corrigir o desbalanceamento de classe, foi empregada uma estratégia para criação de amostras sintéticas. Além disso, foi avaliada a viabilidade do uso de um equipamento portátil (NIR) para tal.

4. METODOLOGIA

4.1 Amostras

Cento e trinta e nove amostras de biodiesel foram fornecidas pelo Laboratório de Ensaios de Combustíveis do Departamento de Química da UFMG (LEC-UFMG). Estas amostras são provenientes do mercado e foram enviadas para análise no LEC no período de 2021/2022. As amostras tiveram sua estabilidade oxidativa determinada pelo método do Rancimat conforme a Resolução ANP Nº 798 DE 01/08/2019 (BRASIL, 2019), sendo 125 aprovadas e 14 reprovadas.

4.2 Análise espectroscópica

4.2.1 ATR-FTIR

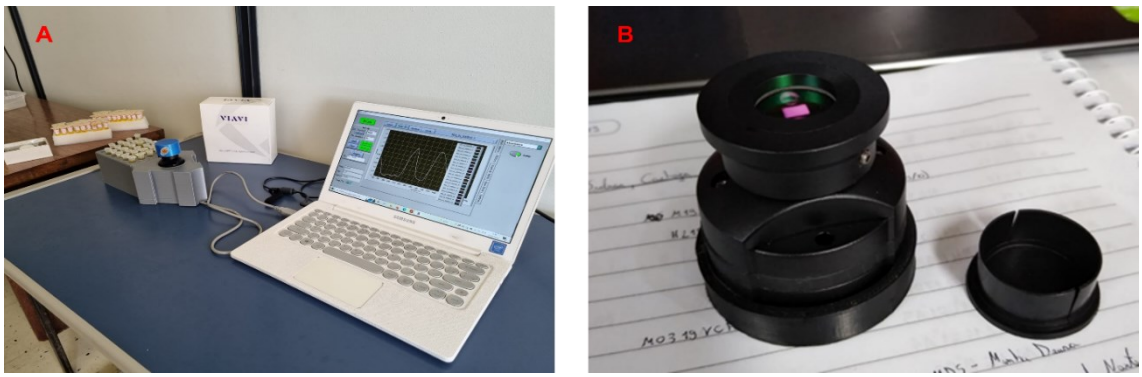
Os espectros no infravermelho médio foram obtidos em um equipamento FTIR, modelo Frontier (Perkin Elmer, Massachusetts, EUA) com um acessório de reflectância total atenuada (ATR) com um cristal de diamante, como pode ser observado na Figura 13. Os espectros foram obtidos na faixa espectral de 4000 – 650 cm^{-1} com uma resolução de 4 cm^{-1} e 32 scans. As medidas foram realizadas com 20 µL de amostras gotejadas no porta amostras em ordem aleatória.

Figura 13 – Equipamento de FTIR

4.2.2 NIR

Os espectros NIR foram obtidos em um equipamento portátil modelo MicroNIR 1700 ES (Viavi Solutions, Arizona, EUA). A faixa espectral utilizada foi de 11000 – 6000 cm^{-1} (900 – 1600 nm), com resolução de 12 nm, tempo de integração 8000 µs e 50 scans. As leituras foram realizadas diretamente do micro tubo e em ordem aleatória. A Figura 14 mostra o equipamento e o suporte utilizado para obtenção dos espectros.

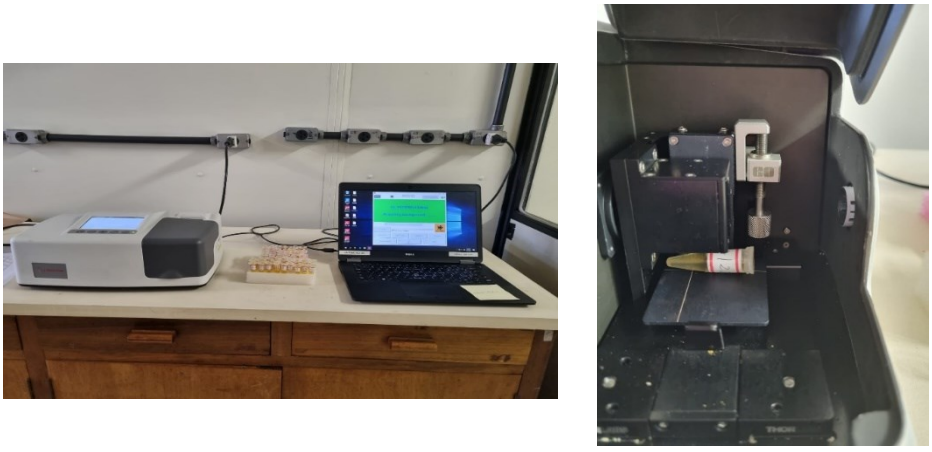
Figura 14 – A - Estação de trabalho do microNIR B- Equipamamento microNIR



4.2.3 Raman

O equipamento para obtenção dos espectros Raman foi um CORA 5700 (Anton Paar, Estíria, Áustria) com um laser em 785 nm, com tempo de exposição de 1s e potência nominal de 200 mW. A faixa espectral utilizada foi de 100 – 2300 cm^{-1} com resolução de 2,25 cm^{-1}. As medidas foram realizadas em ordem aleatória, diretamente do micro tubo como ilustrado na Figura 15.

Figura 15 – Raman CORA 5700



4.2.4 Tratamento de dados

Todo o tratamento de dados foi realizado em software livre, na linguagem de programação Python 3 utilizando software Jupyter notebook (KLUYVER *et al.*, 2016). Três

matrizes de dados foram construídas, uma para cada técnica, todas as matrizes seguiram o mesmo procedimento de tratamento dos dados.

Primeiramente foi feita uma inspeção visual dos espectros obtidos e as amostras consideradas anômalas, devido a ruídos instrumentais, foram descartadas do conjunto de dados. O próximo passo foi o pré-processamento dos dados empregando SNV e derivada com suavização por Savitzky-Golay, métodos comumente empregados em dados espectrais

Após o pré-processamento, o conjunto de dados foi reamostrado pelo método Adasyn (HE, H., BAI, Y., GARCIA, E., & LI, 2008) utilizando a biblioteca Imbalanced-learn (LEMAÎTRE; NOGUEIRA; ARIDAS, 2017). Com a reamostragem dos dados a classe minoritária (não-conforme) foi balanceada com a classe majoritária (conforme). Antes da construção dos modelos, o conjunto de dados foi dividido em conjunto de treinamento, com 70% das amostras, e conjunto teste, com os demais 30%, pelo método de amostragem aleatória estratificada (FORTHOFER; LEE; HERNANDEZ, 2006). Optou-se pelo uso da amostragem aleatória estratificada, pois o subconjunto obtido é um conjunto representativo do conjunto original sem a modificação da proporção de suas classes, além de estar prontamente disponível nas bibliotecas utilizadas.

Em linhas gerais, a amostragem aleatória estratificada pode ser entendida a partir do exemplo: supondo que temos um conjunto de dados C_1 com 100 amostras, 40 pertencentes a classe A_1 e 60 a classe A_2. Necessitamos de um subconjunto C_2 com 50 amostras. Para que as proporções de classes se mantenham as mesmas são necessárias 30 amostras da classe A_1 e 20 da classe A_2. Desta forma retira-se 30 amostras A_1 aleatoriamente sem reposição e 20 amostras de A_2 da mesma maneira.

Antes da construção dos modelos os dados foram centrados na média. Dois métodos de classificação foram empregados para a construção dos modelos de classificação, o LDA e a floresta aleatória ambos utilizando a biblioteca Scikit-learn (BUITINCK *et al.*, 2013). Como discutido na seção 3.4.3 apenas uma variável latente foi utilizada no LDA por se tratar de um problema binário.

Para a construção dos modelos de floresta aleatória foi necessário a otimização dos parâmetros do método tais como: número de árvores, profundidade máxima da árvore, variáveis disponíveis a cada construção e número mínimo de amostras em um nó. Para tal foi realizada uma busca exaustiva por todas as combinações de parâmetros possíveis, os parâmetros testados para os dados de ATR-FTIR, NIR e Raman estão na Tabela 1.

Tabela 1 – Parâmetros empregados na otimização do modelo de floresta aleatória

	Técnica		
	ATR-FTIR	NIR	Raman
Parâmetro			
Número de árvores	De 30 a 150 com incremento de 10	De 50 a 150 com incremento de 10	De 50 a 150 com incremento de 10
Profundidade máxima da árvore	2,3,4,5,7	7,9,15,21	3,5,7,9,15
Número de variáveis disponíveis a cada construção	1950, $\sqrt{1950}$, $\log_2 1950$	88, $\sqrt{88}$, $\log_2 88$	942, $\sqrt{942}$, $\log_2 942$
Número mínimo de amostras por nó	2,3,7,9	2,3,7,9	2,4,6,8

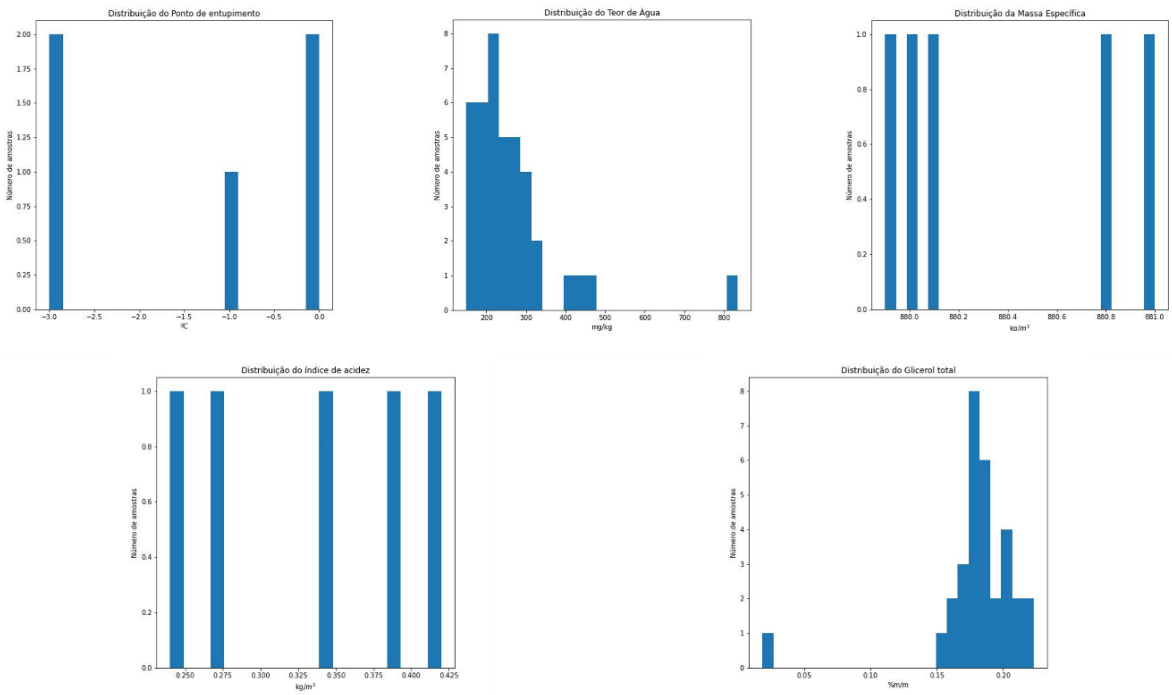
A avaliação do desempenho dos modelos foi feita por meio da matriz de confusão e pela estimativa das figuras de mérito eficiência e MCC, conforme discutido na seção 3.4.2. Modelos sem a etapa de reamostragem também foram construídos da mesma forma descrita acima para comparação do desempenho.

5. RESULTADOS E DISCUSSÕES

5.1 Análise exploratória do conjunto de dados

Devido à natureza das amostras existe uma grande variabilidade entre elas, o que traz uma complexidade extra para a construção de modelos. As amostras possuem origens diferentes, como óleo precursor, presença ou não de antioxidantes, e essas informações não eram conhecidas. A Figura 16 mostra, por exemplo, a distribuição dos parâmetros físico-químicos conhecidos das amostras

Figura 16 – Distribuição de parâmetros Físico-Químicos

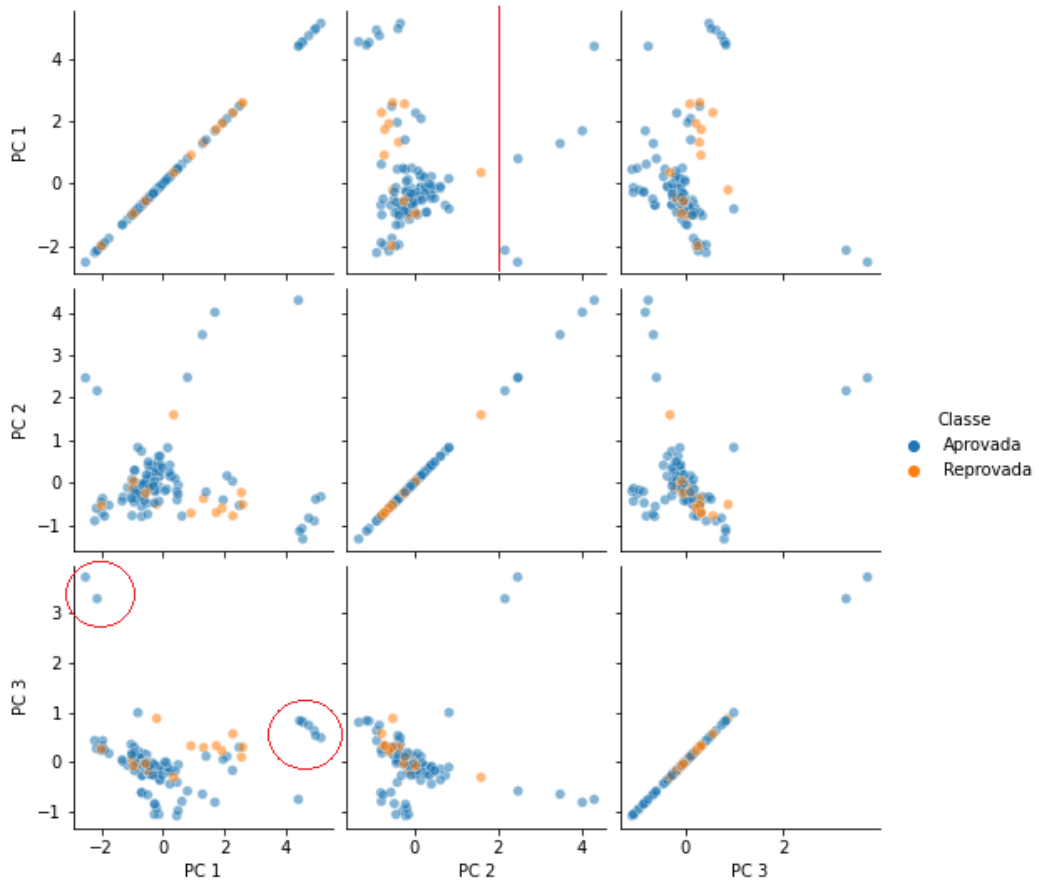


Como podemos observar na Figura 16, o número de amostras com informações físico-químicas é limitado. Mesmo assim nos parâmetros teor de água e glicerol total podemos observar uma grande dispersão nos resultados com faixa analisada indo de 200 – 800 mg/kg para o teor de água e 0,01 – 0,20 para o glicerol total, por exemplo. Os gráficos acima mostram a diversidades das amostras do conjunto de dados empregado na construção dos modelos de classificação.

Na análise exploratória realizada com a PCA, observa-se que 93% da variância dos dados é explicada pelas 3 primeiras PCs, sendo 65% explicada pela PC 1, 19% pela PC 2 e 9%

pela PC 3. Ao observarmos os gráficos de escores, Figura 17, é possível notar uma tendência de agrupamento das amostras não conformes no gráfico de PC 1 x PC 3, entretanto o que fica mais evidente é a grande variabilidade existente nas amostras, seja da classe conforme seja na classe não conforme. No gráfico PC 3 X PC 1 é possível observar algumas amostras bem distintas das demais, sendo duas amostras localizadas no extremo negativo de PC 1, e 10 localizadas no extremo positivo de PC 1. No gráfico de PC 1 X PC 2 é possível observar que sete amostras conformes e uma não conforme se diferenciam das demais, sendo localizadas na porção superior de PC 2. A tendência de agrupamentos distintos não relacionados a classe das amostras ocorre em todas as técnicas utilizadas no trabalho, então optou-se por mostrar apenas a PCA utilizando os dados do ATR-FTIR. As PCAs realizadas com os dados de NIR e Raman estão na seção Anexo.

Figura 17 – Escores da PCA com os dados de MIR.

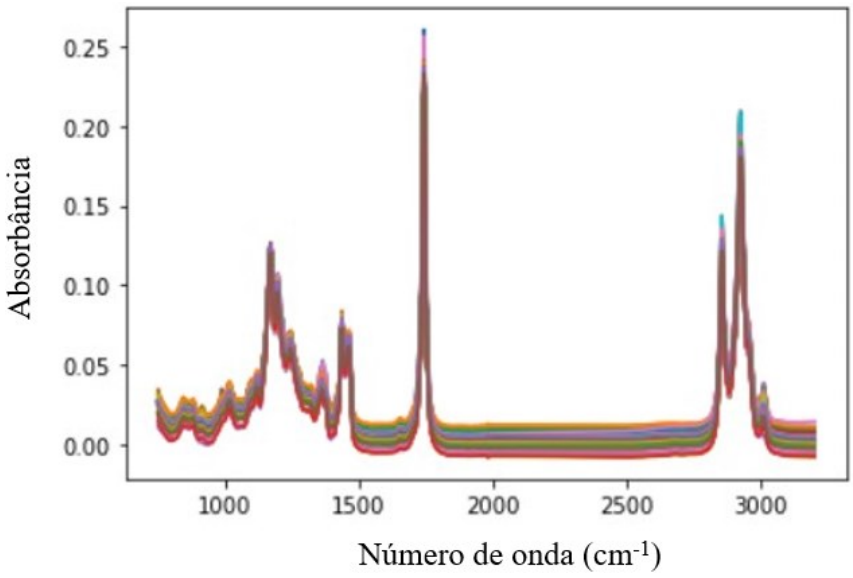


Com o alto grau de desbalanceamento e grande variabilidade das amostras optou-se pelo uso do método de reamostragem Adasyn, descrito na seção 3.4.7 antes da divisão do conjunto de dados em treinamento e teste.

5.2 Modelos de classificação empregando dados de ATR-FTIR

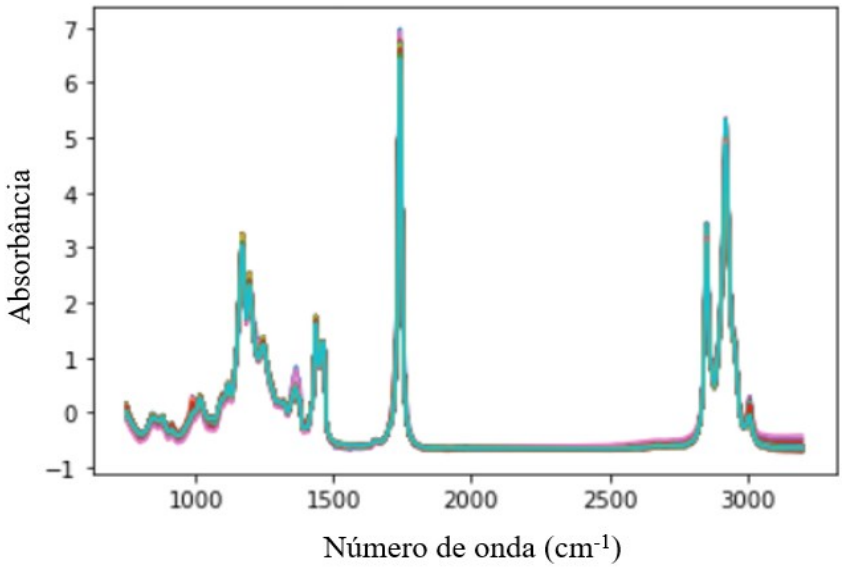
Os espectros foram obtidos como descrito na seção metodologia. Após a inspeção visual, 3 espectros foram retirados do conjunto de dados, pois eles eram claramente diferentes dos demais. A Figura 18 apresenta os espectros brutos, sendo 122 pertencentes a classe das amostras conformes e 14 a classe das amostras não conformes.

Figura 18 – Espectros ATR-FTIR brutos



Pela Figura 18 vemos que a região de 2000 – 2500 cm^{-1} não contém informação química e foi retirada para a construção dos modelos de classificação, assim como as regiões abaixo de 750 cm^{-1} e acima de 3200 cm^{-1}. Após a remoção destas regiões os espectros foram pré-processados, com SNV, e com alisamento Savitzky–Golay, com uma janela de nove pontos e polinômio de segundo grau. Os dados foram reamostrados com o algoritmo Adasyn, totalizando 244 amostras, sendo 122 da classe conforme e 122 da classe não conforme, formando uma matriz 240 x 1950, na qual cada linha representa uma amostra e cada coluna a absorbância para cada número de onda do espectro. A Figura 19 mostra o conjunto de dados reamostrado e pré-processado.

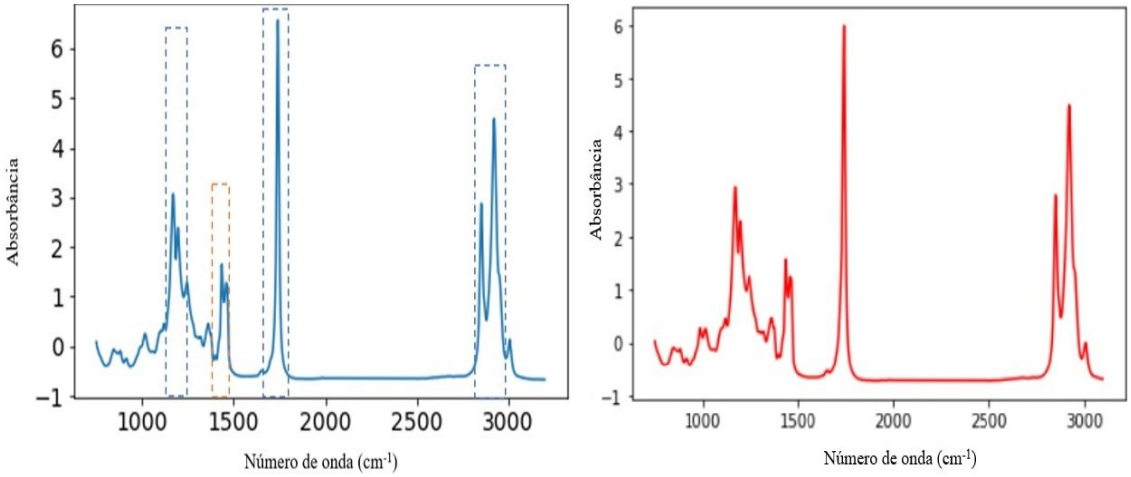
Figura 19 – Espectros ATR-FTIR com a adição de amostras sintéticas e suavizados pelo método SNV e de Savitzky–Golay (janela de 9 pontos)



Como podemos observar na Figura 19 as amostras sintéticas não apresentam grandes diferenças quando comparadas com as originais, uma vez que o espectro reamostrado apresenta o mesmo padrão dos espectros brutos.

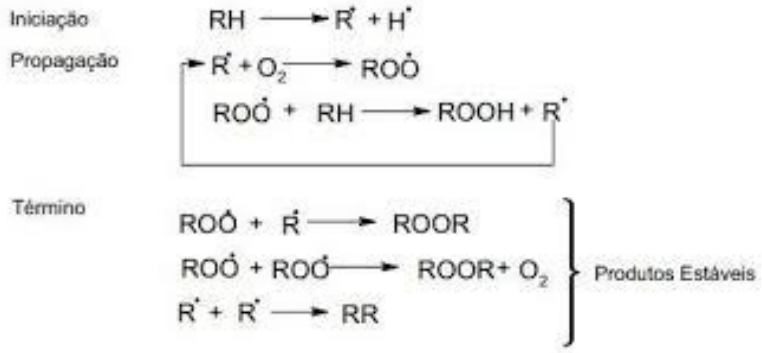
Antes da construção dos modelos foi feita a caracterização espectral de uma amostra de biodiesel, a Figura 20 contém espectros de uma amostra conforme (azul) e não conforme (vermelho) como podemos observar na figura não é possível diferenciar uma amostra conforme de uma não conforme apenas com uma análise visual.

Figura 20 – Espectro ATR-FTIR suavizado de uma amostra conforme em azul e não conforme em vermelho



Antes de caracterização dos espectros propriamente dita é necessária uma breve discussão sobre a oxidação do biodiesel. A oxidação do biodiesel ocorre por mecanismo radicalar que é mecanismo composto por três etapas: a etapa lenta na qual ocorre a formação dos radicais, a etapa de propagação que são reações das moléculas no meio com os radicais formados na etapa lenta e com a liberação de novos radicais e a etapa final que ocorre a formação de moléculas estáveis e o consumo de radicais conforme o esquema apresentado na Figura 21.

Figura 21 - Esquema de reação radicalar



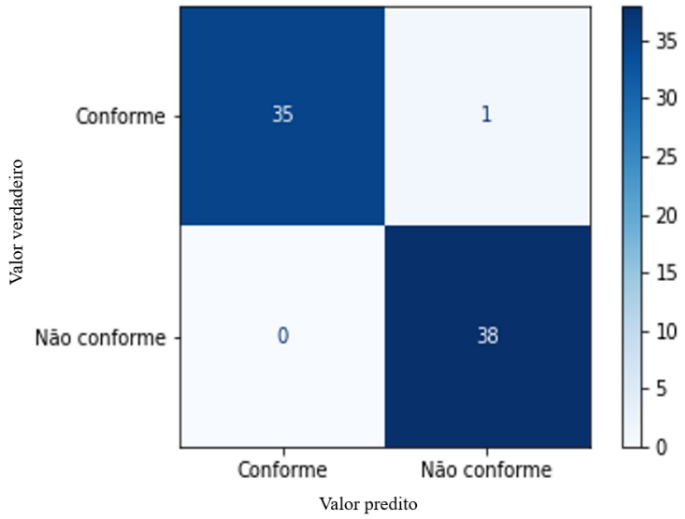
Estudos indicam (YAAKOB *et al.*, 2014) que grupos metileno em posições alílicas e bis-alilícas são mais suscetíveis a perda de um radical H ponto por grupos iniciadores e radicais e os radicais formados nesta etapa podem reagir com O_2 formando radicais peroxila que formam ácidos e radicais R ponto como esquematizados na Figura 21, existem uma grande variedade de produtos conhecidos obtidos da reação de oxidação do biodiesel dentre eles podemos citar: aldeídos, ésteres, ácidos carboxílicos além de dímeros. Com esse conhecimento podemos dar início a caracterização espectral.

As três regiões delimitadas na Figura 20, no espectro da amostra conforme pela linha azul tracejada podem ser atribuídas a: 1100 – 1300 cm^{-1} ao estiramento C-O, 1735 – 1750 cm^{-1} estiramento C=O de ésteres e por fim 2850 – 3000 cm^{-1} a estiramento C-H de alcanos (DE SOUZA; CAJAIBA DA SILVA, 2013). A região delimitada pela linha laranja, 1425 – 1450 cm^{-1} pode ser atribuída a deformação assimétrica de CH_3 (MAHAMUNI; ADEWUYI, 2009). Todas as regiões observadas são condizentes com os produtos formados na oxidação do biodiesel.

5.2.1 LDA

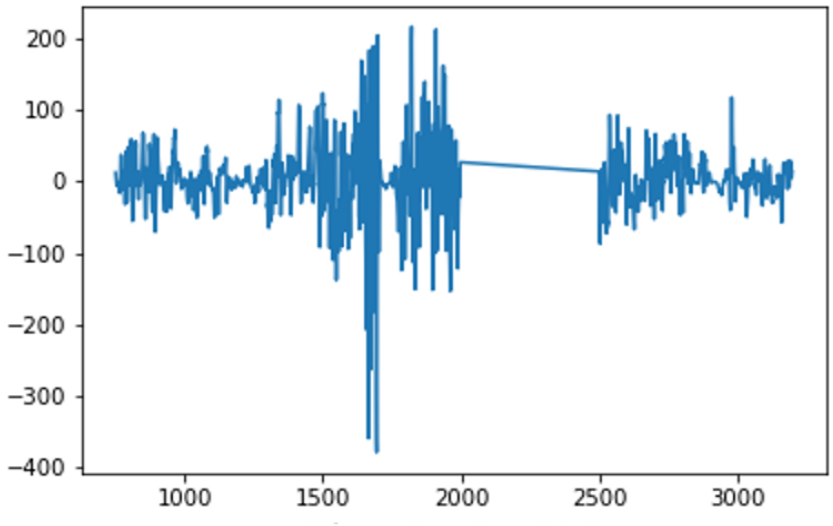
Para a avaliação desempenho do modelo de classificação, a matriz de confusão para o conjunto teste foi construída e está apresentada na Figura 22. A partir dela vemos que o modelo apresenta um ótimo desempenho, com apenas um falso positivo em um total de 74 amostras, com eficiência de 98,6% e um MMC de 0,973.

Figura 22 – Matriz de confusão LDA – dados de ATR-FTIR



O vetor de regressão para o modelo está representado na Figura 23 e, a partir do vetor de regressão, podemos inferir que a região de estiramento C=O 1650 - 1830 cm^{-1} é a mais importante para o modelo, contribuído tanto positivamente quanto negativamente para o vetor de regressão. Isso pode ser atribuído a mudança do número de onda do estiramento C=O de ésteres que se desloca para valores menores (variação de poucos número de onda) que ocorre quando um biodiesel é oxidado, com a formação de produtos de oxidação como aldeídos e cetonas (ZHOU *et al.*, 2017). As regiões próximas a 3000 cm^{-1} e 450 cm^{-1} contribuem positivamente para o modelo, porém em menor escala, e estão relacionadas com o estiramento C-H de alcanos e a deformação assimétrica de CH_3, respectivamente, conforme citado anteriormente.

Figura 23 – Vetor de regressão LDA – dados de ATR-FTIR



Uma abordagem sem reamostragem também foi testada obtendo-se resultados de desempenho inferior ao conjunto reamostrado. Os resultados estão apresentados no Anexo. A eficiência do conjunto teste foi de 92,7%, com 36 verdadeiros positivos, 2 verdadeiros negativos, 1 falso positivo e 2 falsos negativos. Como discutido anteriormente, o parâmetro eficiência não é uma boa métrica para avaliar o desempenho de modelos desbalanceados. O MCC obtido foi de 0,539 indicando um desempenho significativamente menor que o modelo reamostrado.

5.2.2 Floresta aleatória

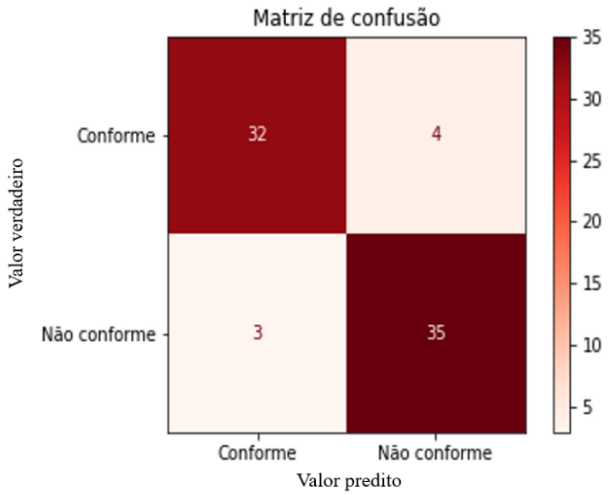
Para este modelo foi necessária uma etapa adicional pois o número de parâmetros a ajustar no modelo, tais como profundidade da árvore, máximo de variáveis disponíveis a cada construção de árvore, número de árvores utilizadas, entre outros é maior que no LDA. Após a busca descrita em metodologia, definiu-se os parâmetros ótimos, descritos na Tabela 2.

Tabela 2 – Parâmetros ótimos para a Floresta Aleatória

Parâmetros	Valores ótimos
Número de árvores	40
Profundidade máxima da árvore	5
Variáveis disponíveis a cada construção	$\sqrt{1950}$
Número mínimo de amostras em um nó	2

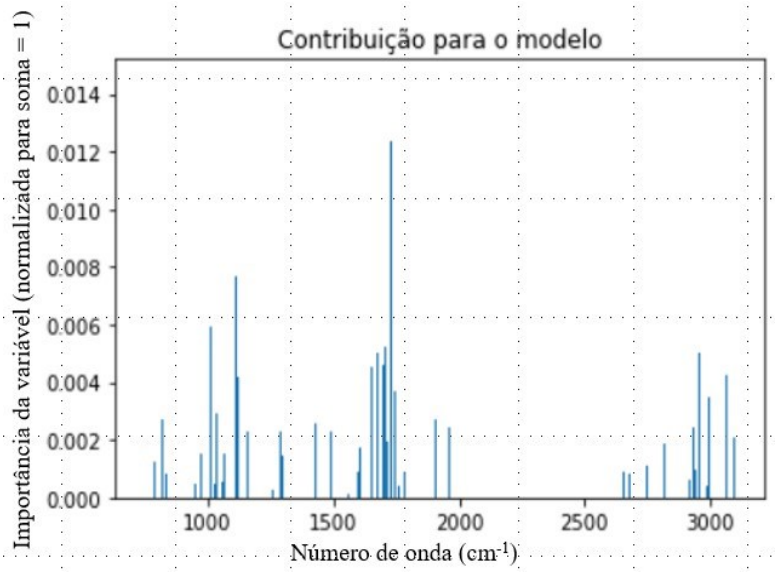
O modelo teve uma eficiência de 100% para o conjunto de treinamento, com todas as amostras classificação corretamente. O conjunto teste foi empregado para avaliação do modelo, a matriz de confusão está representada na Figura 22. A partir dela podemos inferir que o modelo apresenta bom desempenho com 32 verdadeiros positivos e 35 verdadeiros negativos, 4 falsos positivos e 3 falsos negativos com uma eficiência de 90,4%, e um MCC de 0,811.

Figura 24 – Matriz de confusão Floresta aleatória, dados de ATR-FTIR



A Figura 25 representa a importância das variáveis paro o modelo. A partir dela concluímos que as regiões mais importantes para o modelo são: região próxima a 3000 cm^{-1} relacionada ao estiramento C-H de alcanos, a região entre 1650 – 1800 cm^{-1} relacionada ao estiramento C=O e a região de *fingerprint* 1100 – 1400 cm^{-1}. Como discutido anteriormente, essas bandas estão relacionadas como modos vibracionais que sofrem modificação com a oxidação das amostras de biodiesel.

Figura 25 – Importância das variáveis para o modelo de floresta aleatória - dados de ATR-FTIR



Para o modelo de floresta aleatória sem reamostragem os parâmetros ótimos foram: número de árvores 70, profundidade máxima da árvore 2, variáveis disponíveis a cada construção √1950, número mínimo de amostras em um nó 2. Com isto, vemos que o modelo resultante apresenta um comportamento diferente do modelo com reamostragem, possivelmente mais simples, pois apesar do número de árvores ter aumentado em 30 a complexidade de cada árvore diminuiu porque a profundidade máxima diminuiu em 3 a importância das variáveis. A Figura 3A no Anexo, corrobora com essa hipótese pois o número de variável com importância maior que zero é significativamente menor que no modelo com reamostragem. A eficiência no conjunto teste foi de 90,2% com 35 verdadeiros negativos, 2 verdadeiros positivos, 2 falsos positivos e 2 falsos negativos e o MCC foi 0,446.

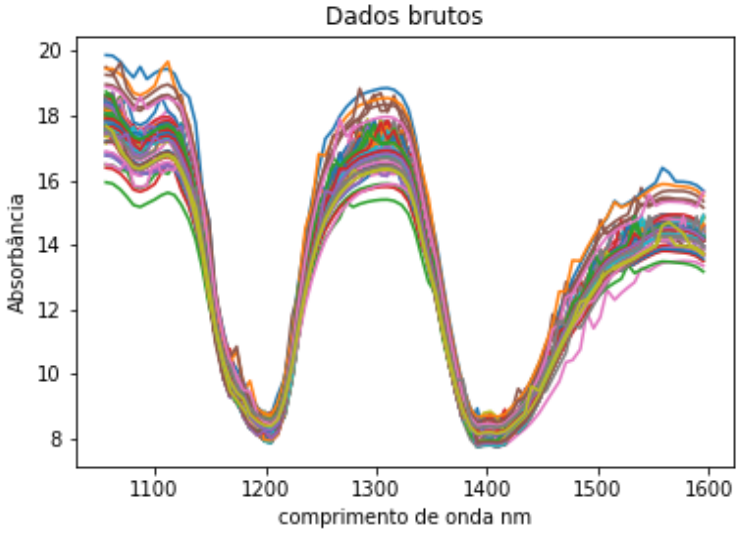
Para os dados de ATR-FTIR a abordagem mais simples, o modelo linear com LDA se mostrou mais adequado, pois obteve um desempenho superior a floresta aleatória.

5.3 Modelos de classificação empregando dados de NIR

Os espectros foram obtidos com um equipamento portátil conforme descrito na seção 5.2.1. Os espectros foram pré-processados com SNV, segunda derivada, alisamento Savitzky–Golay, com uma janela de 19 pontos e polinômio de segundo grau, e posteriormente reamostrados com o Adasyn resultando em uma matriz 249 x 88, com 125 amostras da classe

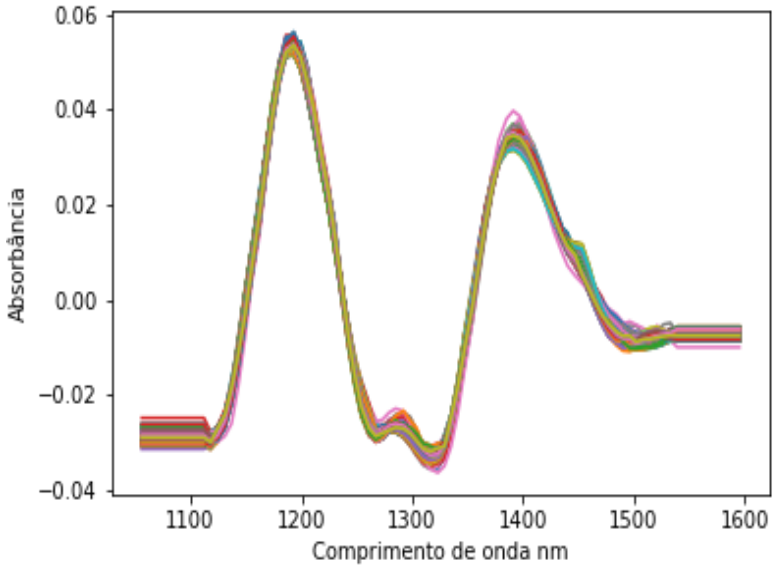
conforme e 124 amostras da classe não conforme. Essa diferença no número de amostras entre as classes pode ocorrer caso o algoritmo não encontre mais amostras na região duvidosa. A Figura 26 mostra os espectros obtidos.

Figura 26 – Dados brutos NIR



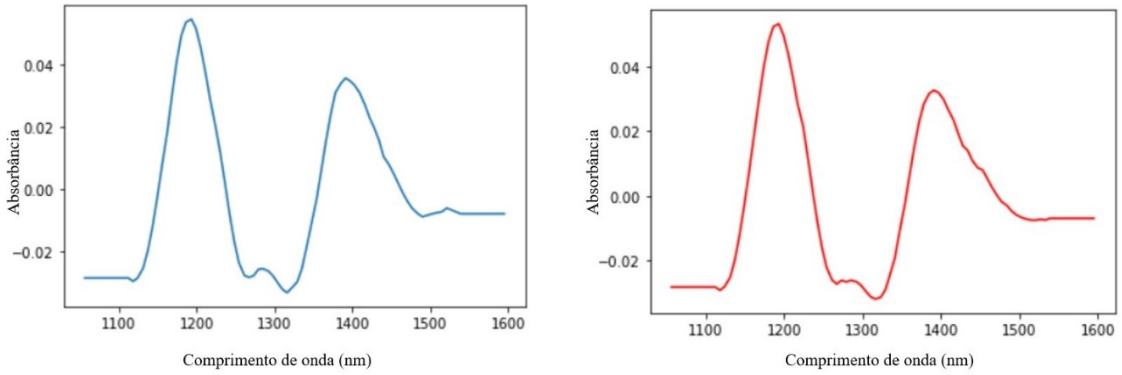
Podemos observar a presença de ruído nos espectros, provavelmente devido às limitações instrumentais. Apenas a região entre 1050 – 1600 nm foi considerada, eliminando os extremos do espectros para a construção dos modelos. A Figura 27 mostra os espectros pré-processados e reamostrados.

Figura 27 – Dados NIR suavizados com SNV e Savitzky–Golay, com 2ª derivada e janela de 19 pontos



O espectro de uma amostra conforme em azul e de uma amostra não conforme em vermelho pré-processados como descrito anteriormente estão representados na figura 28.

Figura 28 – Espectros NIR pré-processados de uma amostra de biodiesel conforme (azul) e uma não conforme (vermelho)

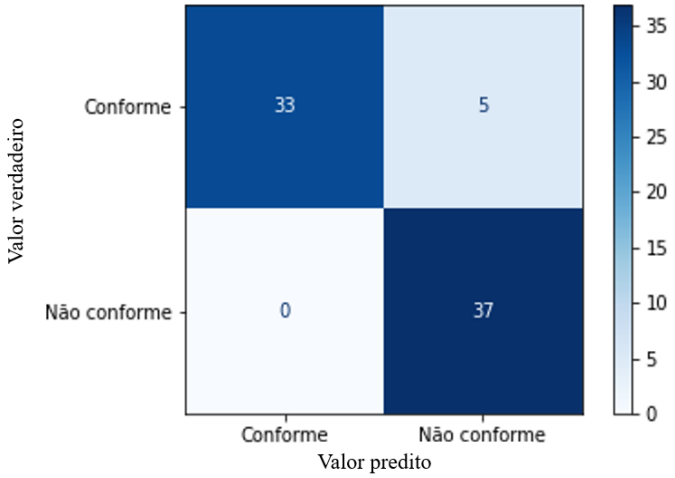


Sabe-se que a região entre 1100 – 1600 nm é a região de segundo sobretom de CH, CH_2, CH_3 e o primeiro sobretom de R-OH (SILVA *et al.*, 2012). Todas são transições citadas são condizentes com os produtos formados durante a oxidação do biodiesel, conforme discussão anterior.

5.3.1 LDA

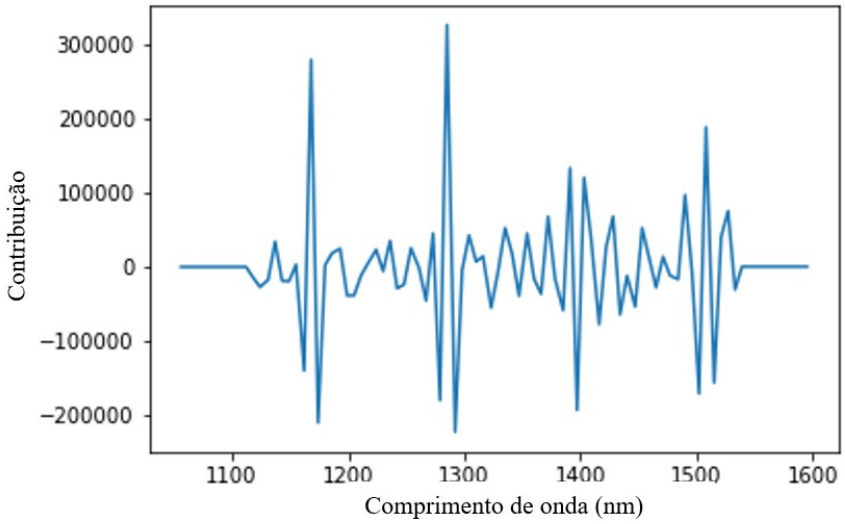
O modelo apresentou um desempenho perfeito para o conjunto de treinamento, com todas as amostras classificadas corretamente. A matriz de confusão para o conjunto teste, usada para a avaliação do modelo, está representada na Figura 29. Com base nessas informações podemos observar que o modelo apresenta um bom desempenho com 33 verdadeiros positivos, 5 falsos positivos, 37 verdadeiros negativos e nenhum falso negativo com uma eficiência de 93,3% e um MCC de 0,875.

Figura 29 – Matriz de confusão LDA – dados de NIR



Ao analisarmos os coeficientes do vetor de regressão para este modelo, Figura 30, notamos quatro regiões com maior importância para o modelo, com bandas em torno de 1200, 1300, 1400 e 1500 nm.

Figura 30 – Vetor de regressão LDA – dados de NIR



Uma abordagem sem a reamostragem dos dados também foi testada, resultando em 35 VP, 3 FN, 0 VN e 4 FP com eficiência de 83,3% e um MCC de -0,090. Pelo coeficiente de regressão, apresentado no Anexo, notamos que as regiões mais importantes não mudam, mas a escala do coeficiente muda, bem como o padrão das bandas importantes nessas regiões. Por exemplo, a região de 1400 nm no modelo com reamostragem apresenta uma banda contribuindo negativamente de maneira mais intensa, já no modelo sem reamostragem duas outras bandas nessa região podem ser notadas contribuindo positivamente. Sem a reamostragem o modelo é

incapaz de capturar a classe minoritária e classifica todas as amostras como conformes sendo inadequado para o problema.

5.3.2 Floresta aleatória

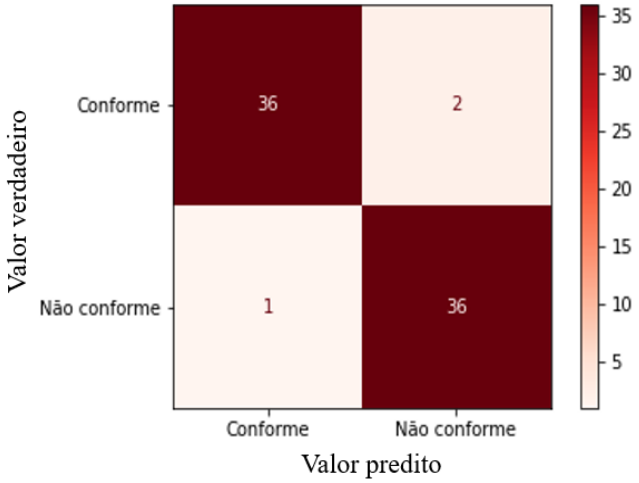
Após a busca descrita na seção de Metodologia definiu-se os parâmetros ótimos, descritos na Tabela 3 para a construção do modelo.

Tabela 3 – Parâmetros ótimos para a Floresta Aleatória – dados de NIR

Parâmetros	Possíveis valores
Número de árvores	50
Profundidade máxima da árvore	9
Variáveis disponíveis a cada construção	$\log_2 88$
Número mínimo de amostras em um nó	2

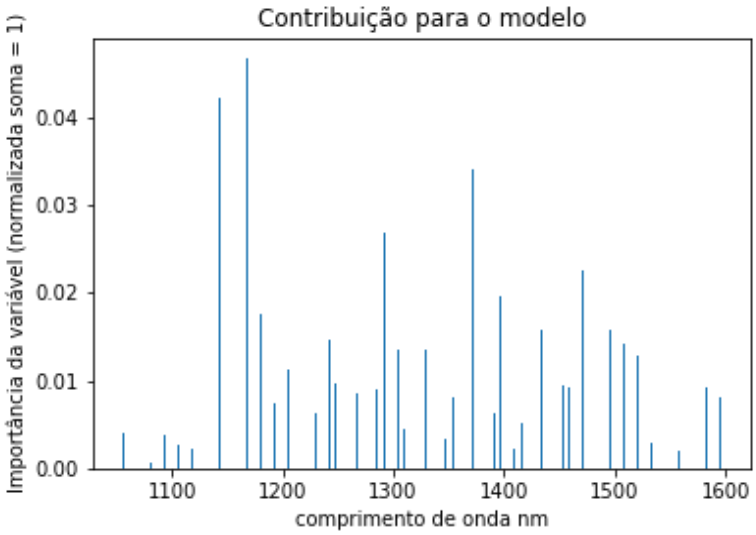
O modelo de floresta aleatória com a reamostragem dos dados obteve uma eficiência de 97,3% e um MCC de 0,920 no conjunto teste, com apenas 1 falso positivo e 2 falsos negativos, como demonstrado na matriz de confusão da Figura 31, sendo um modelo adequado para a classificação das amostras de biodiesel.

Figura 31 – Matriz de confusão floresta aleatória – NIR



A importância das variáveis para o modelo está representada na Figura 32. A partir dela notamos um comportamento similar ao observado nos coeficientes do vetor de regressão do LDA com quatro regiões mais importantes, a região entre 1050 – 1200 nm, com a maior contribuição para o modelo, e as regiões em torno de 1300 e 1400 nm e 1500 nm.

Figura 32 – Contribuição das variáveis para a floresta aleatória - dados de NIR



Ao aplicarmos o modelo de floresta aleatória nos dados de NIR, sem reamostragem, os parâmetros ótimos são: número de árvores 50, profundidade máxima da árvore 7, variáveis disponíveis a cada construção $\sqrt{88}$, número mínimo de amostras em um nó 2. Assim como no LDA, o desempenho do modelo cai substancialmente, todas as amostras foram classificadas

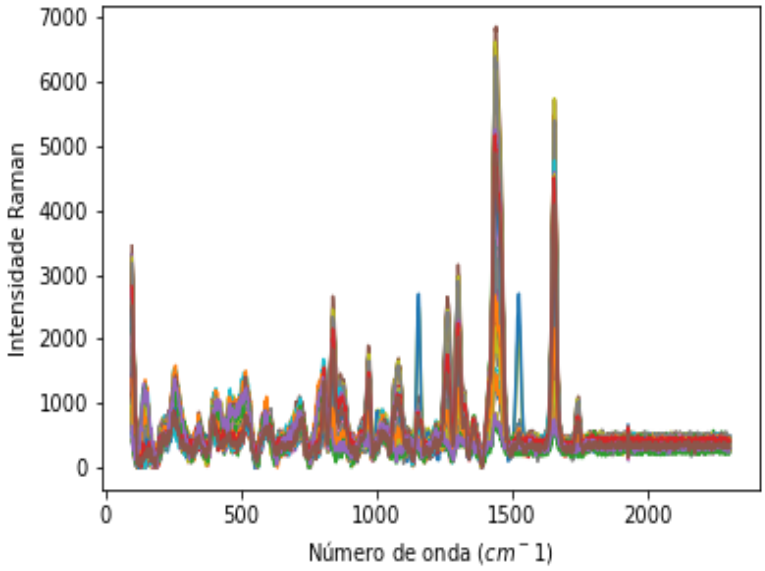
como conformes atingindo uma eficiência de 90,4% e um MCC de 0. Ao observarmos a importância das variáveis para o modelo, Figura 4A no Anexo, notamos que a importância das em torno de 1400 nm aumenta consideravelmente.

Com os dados de NIR a reamostragem se mostrou necessária, pois os modelos sem a reamostragem não foram capazes de capturar a variância dos dados experimentais apesar dos altos valores de eficiência apresentados. Com a reamostragem, os MCC de ambos foi acima de 0,9, este aumento no desempenho está associado a modificações nos coeficientes dos vetores de regressão do modelo LDA, ou a modificação da importância das variáveis para a floresta aleatória decorrentes do balanceamento das classes.

5.4 Construção dos modelos com dados de Raman

Os espectros Raman foram obtidos conforme descritos na seção Metodologia. Os dados brutos estão apresentados na Figura 33.

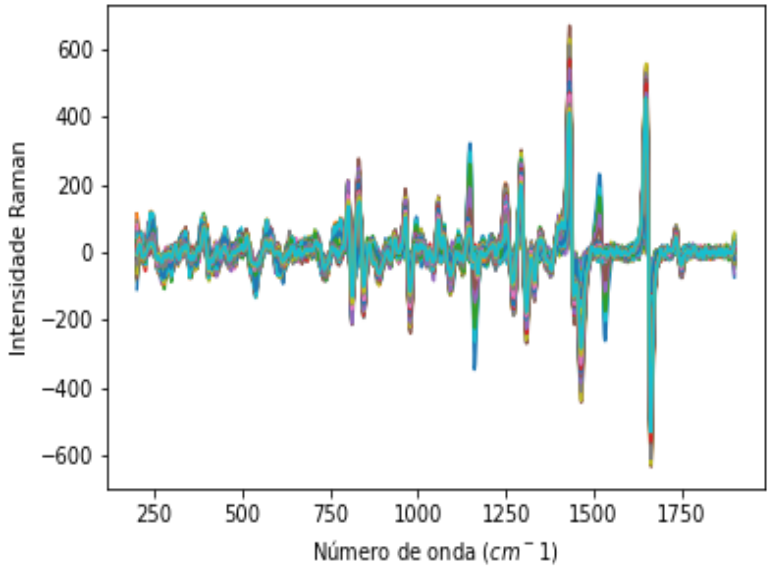
Figura 33 – Espectros brutos Raman



Os espectros foram limitados de 200 – 1900 cm^{-1} para eliminar a região de ruído instrumental. Em sequência foi realizada a primeira derivada seguida de suavização Savitsky-Golay com janela de nove pontos e polinômio de segundo grau. Após o pré-processamento foi realizada a reamostragem resultando em uma matriz 240 X 942, sendo que nessa matriz 122

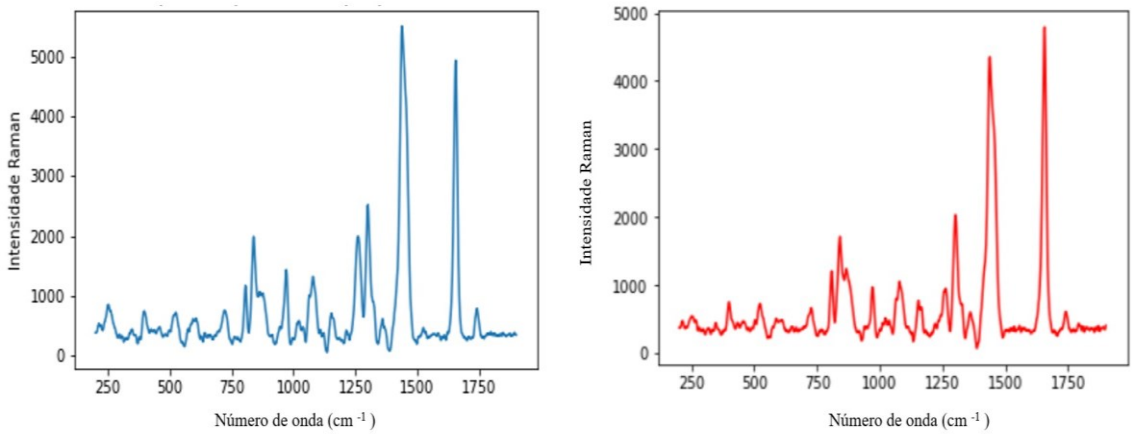
amostras são da classe conforme e 118 amostras da classe não conforme. A diferença do número de amostras entre as classes após o desbalanceamento ocorreu pelo mesmo motivo que nos dados de NIR. A Figura 34 mostra os espectros reamostrados e pré-processados.

Figura 34 – Dados reamostrados e pré-processados com SNV e Savitzky–Golay, com primeira derivada e janela de 9 pontos - dados Raman



Assim como nos outros modelos as amostras sintéticas não diferem das demais, indicando que a reamostragem foi adequada. A Figura 35 traz os espectros pré-processados, mas sem a primeira derivada de uma amostra de biodiesel conforme (azul) e não conforme (vermelho), e serão utilizados na caracterização espectral.

Figura 35 – Espectro Raman pré-processado sem derivada de uma amostra conforme em azul e não conforme em vermelho

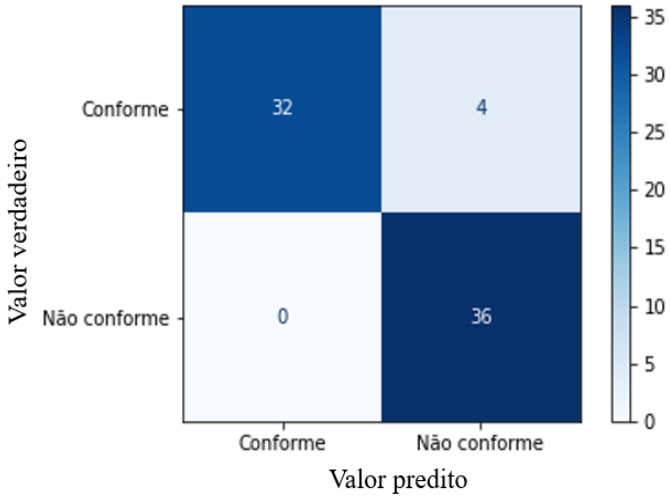


A banda na região de 1700 – 1750 cm^{-1} pode ser atribuída ao estiramento C=O, a banda na região de 1600 – 1680 cm^{-1} pode ser atribuída ao estiramento C=C esperada de ser encontrada devido as insaturações presentes no biodiesel, a banda na região de 1400 – 1500 cm^{-1} pode ser atribuída como deformações CH_2, CH_3. As bandas na região de 1245 – 1277 cm^{-1} podem ser atribuídas a deformações C-H do tipo tesoura. A banda na região de 900 – 1000 cm^{-1} é atribuída a deformações C-H do tipo torção. Na região de 800 – 900 cm^{-1} também podem ser observadas bandas de estiramento C-O (MIRANDA *et al.*, 2014). Apesar de pequenas diferenças entre os espectros Raman de uma amostra conforme e não conforme, não é possível classificar uma amostra de biodiesel quanto a sua estabilidade oxidativa apenas com a análise visual do espectro.

5.4.1 LDA

O modelo construído com os dados de reamostragem obteve uma eficiência de 94,4% e um MCC de 0,894 para o conjunto teste com podemos inferir pela matriz de confusão na Figura 36.

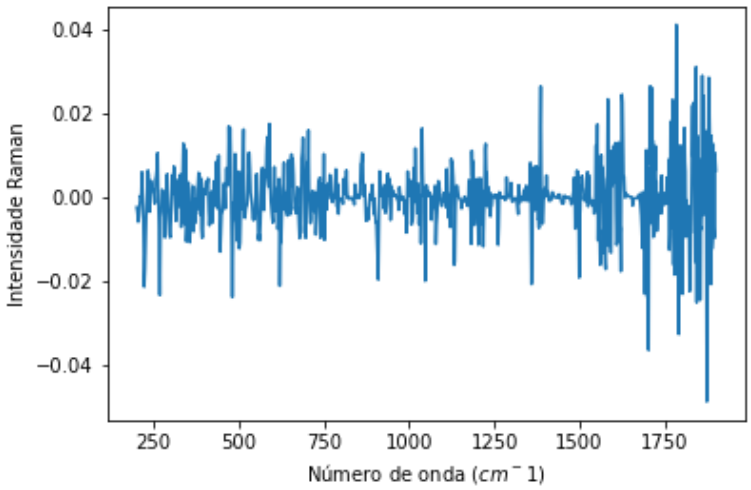
Figura 36 – Matriz de confusão para o LDA com reamostragem – dados Raman



Ao observarmos os coeficientes do vetor de regressão na Figura 37, podemos observar que as bandas na região de 1750 cm^{-1} e 1650 cm^{-1} são as mais importantes para o modelo. Essa

região pode ser atribuída a modos vibracionais C=O e C=C, e estão diretamente relacionadas com as mudanças que ocorrem na composição do biodiesel após processos de oxidação.

Figura 37 – Coeficiente do vetor de regressão LDA- dados de Raman



Com os dados sem reamostragem o modelo não foi capaz de capturar a variabilidade das amostras da classe não conforme e classificou todas as amostras como pertencentes a classe conforme, atingindo uma eficiência de 92,7% e um MCC igual a 0. Na figura 3A do Anexo podemos observar que a importância das variáveis diminui para aproximadamente a metade quando comparada ao modelo com reamostragem.

5.4.2 Floresta aleatória

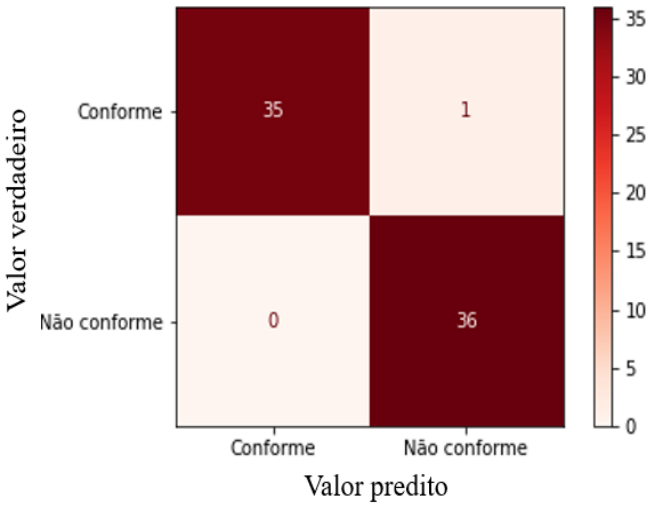
A busca pelos melhores parâmetros foi realizada da mesma maneira que para os dados das outras técnicas espectroscópicas. Os parâmetros ótimos estão descritos na Tabela 4.

Tabela 4 – Parâmetros ótimos para a floresta aleatória - dados Raman

Parâmetros	Parâmetros ótimos
Número de árvores	60
Profundidade máxima da árvore	5
Variáveis disponíveis a cada construção	$\sqrt[2]{942}$
Número mínimo de amostras em um nó	2

A avaliação do modelo foi feita a partir do conjunto de teste. O modelo obteve eficiência de 98,6% e um MCC de 0,972 para o conjunto teste, como podemos calcular a partir da matriz de confusão, na Figura 38. Apenas um falso negativo foi observado.

Figura 38 – Matriz de confusão floresta aleatória - dados Raman



Ao observarmos a importância das variáveis para o modelo, Figura 39, observamos a banda mais importante na região próxima 1400 cm^{-1}, que é a região de deformações da ligação CH, como citado anteriormente. Além disso as demais regiões citadas anteriormente apresentam importâncias parecidas, indicando que o modelo consegue capturar bem a informação química presente no espectro.

Figura 39 - Importância das variáveis para a floresta aleatória - dados Raman

Uma abordagem sem reamostragem também foi testada, no entanto, o modelo não apresentou desempenho adequado, pois classifica todas as amostras como conformes com uma eficiência de 92,7% e um MCC de 0. Esse resultado é comum para classes desbalanceadas, em que o modelo apresenta capacidade preditiva apenas apara a classe majoritária.

5.5 Considerações dos Modelos de classificação

A partir dos resultados apresentados é possível afirmar que todas as três técnicas espectroscópicas vibracionais podem ser utilizadas para a classificação do biodiesel, segundo a estabilidade oxidativa, pois tanto modelos lineares quanto não lineares obtiveram bons resultados quando associados a reamostragem. A figura de mérito eficiência não se mostrou muito útil quando o desbalanceamento de classes é grande, como era o caso do conjunto de dados, pois mesmo quando todas as amostras eram atribuídas a classe majoritária seu valor era superior a 90% dando a impressão de que o modelo era adequado. Portanto as discussões feitas aqui serão em termos do coeficiente de correlação de Matthews. A Tabela 5 resume os resultados obtidos para os modelos construídos com e sem reamostragem.

Tabela 5 – Comparação dos modelos de classificação para as diferentes técnicas

Técnicas	**Modelos**			
	LDA		**Floresta Aleatória**	
	Reamostragem		**Reamostragem**	
	Não	**Sim**	**Não**	**Sim**
MIR	0,539	0,973	0,446	0,811
NIR	-0,09	0,875	0	0,920
Raman	0	0,894	0	0,972

A reamostragem das amostras de todo o conjunto de dados se mostrou uma ferramenta adequada para compensar a alta variabilidade existente nas amostras, associadas ao desbalanceamento das classes. Todos os modelos construídos com os dados reamostrados apresentaram melhor performance.

Os modelos construídos a partir dos dados de infravermelho médio, apresentaram coeficientes de correlação de Matthews próximos a 0,9. O menor valor encontrado, 0,811, foi no conjunto de dados de FTIR para o modelo da floresta aleatória. Neste caso, o modelo mais simples, LDA, se mostrou mais adequado para classificar as amostras de biodiesel em conforme, de acordo com a estabilidade oxidativa. Entre as técnicas espectroscópicas, a ATR-FTIR foi a única nas quais os modelos sem reamostragem apresentaram melhor desempenho.

O uso de equipamentos portáteis é uma tendência na área de espectroscopia vibracional, permitindo análises rápidas e *in-situ*, e abrindo novas possibilidades, como a incorporação desses dispositivos em smartphones em um futuro próximo (Elton, 2018). Neste trabalho, os dados de NIR foram obtidos em um equipamento portátil, com a aquisição dos espectros direto no recipiente, sem nenhum contato com o analista. O melhor modelo obtido foi empregando a floresta aleatória com reamostragem, a eficiência obtida foi de 97% e MCC de 0,920.

A técnica de espectroscopia Raman também forneceu ótimos resultados quando associada a técnica de reamostragem Adasyn. Os valores de MCC foram 0,894 para o LDA e 0,972 para a floresta aleatória. Para os dados sem a reamostragem, a capacidade preditiva do modelo diminuiu para a classe minoritária e o MCC foi igual a 0. Os dados de espectroscopia Raman foram obtidos direto no recipiente da amostra e as medidas foram realizadas em torno de 30s.

O modelo com LDA e dados de infravermelho médio e o modelo de floresta aleatória com os dados de espectroscopia Raman tiveram o melhor desempenho com os dados

reamostrados. Seguindo o princípio da parcimônia, métodos mais simples devem ser testados e usados, com essa consideração, o modelo LDA com os dados de infravermelho médio apresenta a vantagem da simplicidade contra o alto custo computacional durante a etapa de otimização dos parâmetros da floresta aleatória.

6. CONSIDERAÇÕES FINAIS

Neste trabalho, modelos de classificação foram construídos a partir de dados de espectroscopia vibracional e ferramentas de *machine learning* para classificar amostras de biodiesel B100 em conforme e não conforme em relação a estabilidade oxidativa. Os melhores modelos apresentaram eficiência e MCC acima de 0,9 para os dados reamostrados.

A estratégia de reamostragem foi empregada para o balanceamento das classes ajustando o número de amostras da classe minoritária, melhorando o desempenho da capacidade preditiva dos modelos. O modelo linear, LDA, apresentou melhor resultado para os dados de infravermelho médio, com eficiência de 98,6% e MCC igual a 0,973. Já os modelos com os dados de NIR e Raman empregando o método floresta aleatória alcançaram melhores desempenho. Para o NIR, a eficiência foi de 97% e o MCC de 0,920, e para os dados de Raman, foi obtida uma eficiência de classificação de 98,6% e MCCC de 0,972.

A caracterização dos modelos foi realizada avaliando a importância das variáveis espectrais para a discriminação das amostras. Regiões relacionadas ao estiramento C=C e C=O apresentaram importância nos modelos e são condizentes com o parâmetro avaliado.

Por fim, a utilização de técnicas espectroscópicas associadas à métodos de *machine learning* para a classificação de biodiesel segundo as normas da ANP para a estabilidade oxidativa é promissora, permitindo a inspeção de modo direto das amostras de biodiesel B100, sem nenhum preparo, de forma rápida e com detecção *in loco* com o uso de equipamentos portáteis, disponíveis comercialmente para as três técnicas empregadas neste trabalho.

REFERÊNCIAS

ALMEIDA, Mariana R. Espectroscopia Raman e quimiometria como ferramentas analíticas para química forense e paleontologia. Tese (Doutorado em Química) – UNICAMP, Campinas, 2015.

AMIGO, José Manuel. Data mining, machine learning, deep learning, chemometrics: Definitions, common points and trends (Spoiler Alert: VALIDATE your models!). **Brazilian Journal of Analytical Chemistry**, [*s. l.*], v. 8, n. 32, p. 22–38, 2021.

BATISTA, Gustavo E. A. P. A.; PRATI, Ronaldo C.; MONARD, Maria Carolina. A study of the behavior of several methods for balancing machine learning training data. **ACM SIGKDD Explorations Newsletter**, [*s. l.*], v. 6, n. 1, p. 20–29, 2004. Disponível em: https://dl.acm.org/doi/10.1145/1007730.1007735. Acesso em: 10/10/2022.

BLANCHARD, Gary J; BROWN, Steven D. Book Reviews: Introduction to Multivariate Statistical Analysis in Chemometrics. **Applied Spectroscopy**, [*s. l.*], v. 64, n. 4, p. 112A-112A, 2010. Disponível em: https://doi.org/10.1366/000370210791114185. Acesso em: 05/03/2021.

BREIMAN, Leo. Random forests. **Machine Learning**, [*s. l.*], v. 45, p. 5–32, 2001. Disponível em: https://link.springer.com/article/10.1023%2FA%3A1010933404324#citeas.
BUITINCK, Lars *et al.* {API} design for machine learning software: experiences from the scikit-learn project. *In*: , 2013. **ECML PKDD Workshop: Languages for Data Mining and Machine Learning**. [*S. l.: s. n.*], 2013. p. 108–122.

BUKKARAPU, Kiran Raj; KRISHNASAMY, Anand. Predicting engine fuel properties of biodiesel and biodiesel-diesel blends using spectroscopy based approach. **Fuel Processing Technology**, [*s. l.*], v. 230, n. February, p. 107227, 2022. Disponível em: https://doi.org/10.1016/j.fuproc.2022.107227. Acesso em: 02/06/2020.

CORREIA, Radigya M. *et al.* Portable near infrared spectroscopy applied to fuel quality control. **Talanta**, [*s. l.*], v. 176, p. 26–33, 2018. Disponível em: http://dx.doi.org/10.1016/j.talanta.2017.07.094. Acesso em: 10/10/2022.

DE LIRA, Liliana Fátima Bezerra *et al.* Infrared spectroscopy and multivariate calibration to monitor stability quality parameters of biodiesel. **Microchemical Journal**, [*s. l.*], v. 96, n. 1, p. 126–131, 2010. Disponível em: http://dx.doi.org/10.1016/j.microc.2010.02.014. Acesso em: 30/07/2020.

DE SOUZA, Adriana Velloso A.; CAJAIBA DA SILVA, João F. Biodiesel synthesis evaluated by using real-time ATR-FTIR. **Organic Process Research and Development**, [*s. l.*], v. 17, n. 1, p. 127–132, 2013.

DO AMARAL, Bruna Elói; DE REZENDE, Daniel Bastos; PASA, Vânya Márcia Duarte. Aging and stability evaluation of diesel/ biodiesel blends stored in amber polyethylene bottles under different humidity conditions. **Fuel**, [*s. l.*], v. 279, n. March, p. 118289, 2020. Disponível em: https://doi.org/10.1016/j.fuel.2020.118289. Acesso em: 19/01/2022.

DUNN, Kevin G. Process Improvement using Data. [*s. l.*], n. 294-34b8, p. 381, 2014. Disponível em: learnche.mcmaster.ca/pid.

ELREEDY, Dina; ATIYA, Amir F. A Comprehensive Analysis of Synthetic Minority Oversampling Technique (SMOTE) for handling class imbalance. **Information Sciences**, [*s. l.*], v. 505, p. 32–64, 2019. Disponível em: https://doi.org/10.1016/j.ins.2019.07.070. Acesso em: 09/12/2021.

FERREIRA, Márcia M C. **Quimiometria: conceitos, métodos e aplicações**. [*S. l.*]: Editora da Unicamp, 2015. *E-book*. Disponível em: https://books.google.com.br/books?id=a5OnDwAAQBAJ. Acesso em: 05/08/2022.

FORTHOFER, Ronald N.; LEE, Eun Sul; HERNANDEZ, Mike. **Biostatistics: A Guide to Design, Analysis and Discovery**. [*S. l.: s. n.*], 2006.

GUO, Gongde *et al.* KNN model-based approach in classification. **Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)**, [*s. l.*], v. 2888, p. 986–996, 2003.

HE, Haibo, BAI, Yang, GARCIA, Edwardo, & LI, Shutao. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In IEEE International Joint Conference on Neural Networks, 2008. **IJCNN 2008.(IEEE World Congress on Computational Intelligence) (pp. 1322– 1328)**, [*s. l.*], n. 3, p. 1322– 1328, 2008.

HESS, Christian. New advances in using Raman spectroscopy for the characterization of catalysts and catalytic reactions. **Chemical Society Reviews**, [*s. l.*], v. 50, n. 5, p. 3519–3564, 2021.

HIND, Andrew R.; BHARGAVA, Suresh K.; MCKINNON, Anthony. At the solid/liquid interface: FTIR/ATR - The tool of choice. **Advances in Colloid and Interface Science**, [*s. l.*], v. 93, n. 1–3, p. 91–114, 2001.

KARAVALAKIS, George; STOURNAS, Stamoulis; KARONIS, Dimitrios. Evaluation of the oxidation stability of diesel/biodiesel blends. **Fuel**, [*s. l.*], v. 89, n. 9, p. 2483–2489, 2010. Disponível em: http://dx.doi.org/10.1016/j.fuel.2010.03.041. Acesso em: 12/03/2022.

KELLEHER, Jhon D; NAMEE, Brian M; D'ARCY, Aoife. **Fundamentals of Machine Learning for Predictive Data Analytics, second edition: Algorithms, Worked Examples, and Case Studies**. [*S. l.*]: MIT Press, 2020. *E-book*. Disponível em: https://books.google.com.br/books?id=1Iv-DwAAQBAJ. Acesso em: 25/08/2020.

KLUYVER, Thomas *et al.* Jupyter Notebooks -- a publishing format for reproducible computational workflows. *In*: , 2016. (F Loizides & B Schmidt, Org.)**Positioning and Power in Academic Publishing: Players, Agents and Agendas**. [*S. l.: s. n.*], 2016. p. 87–90.

KOVÁCS, György. An empirical comparison and evaluation of minority oversampling techniques on a large number of imbalanced datasets. **Applied Soft Computing**, [*s. l.*], v. 83, p. 105662, 2019. Disponível em: https://doi.org/10.1016/j.asoc.2019.105662. Acesso em: 22/09/2020.

KUMAR, Niraj. Oxidative stability of biodiesel: Causes, effects and prevention. **Fuel**, [*s. l.*], v. 190, p. 328–350, 2017. Disponível em: http://dx.doi.org/10.1016/j.fuel.2016.11.001.
LEMAÎTRE, Guillaume; NOGUEIRA, Fernando; ARIDAS, Christos K. Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning. **Journal of Machine Learning Research**, [*s. l.*], v. 18, n. 17, p. 1–5, 2017. Disponível em: http://jmlr.org/papers/v18/16-365.html. Acesso em: 16/08/2022.

LEUNG, Dennis Y.C.; WU, Xuan; LEUNG, M. K.H. A review on biodiesel production using catalyzed transesterification. **Applied Energy**, [*s. l.*], v. 87, n. 4, p. 1083–1095, 2010. Disponível em: http://dx.doi.org/10.1016/j.apenergy.2009.10.006. Acesso em: 20/08/2022.

LOUPPE, Gilles. Understanding Random Forests: From Theory to Practice. [*s. l.*], n. July, 2014. Disponível em: http://arxiv.org/abs/1407.7502. Acesso em: 21/08/2021.

MAHAMUNI, Naresh N.; ADEWUYI, Yusuf G. Fourier transform infrared spectroscopy (FTIR) method to monitor soy biodiesel and soybean oil in transesterification reactions, petrodiesel- biodiesel blends, and blend adulteration with soy oil. **Energy and Fuels**, [*s. l.*], v. 23, n. 7, p. 3773–3782, 2009.

MCCORMICK, Robert L.; WESTBROOK, Steven R. Storage stability of biodiesel and biodiesel blends. **Energy and Fuels**, [*s. l.*], v. 24, n. 1, p. 690–698, 2010.

MIRANDA, Alisson M. *et al.* Line shape analysis of the Raman spectra from pure and mixed biofuels esters compounds. **Fuel**, [*s. l.*], v. 115, n. 0016, p. 118–125, 2014. Disponível em: http://dx.doi.org/10.1016/j.fuel.2013.06.038. Acesso em: 03/03/2022.

MÜLLER, Andreas. **Introduction to Machine Learning with Python: A Guide for Beginners in Data Science**. [*S. l.: s. n.*], 2018.

MURTA VALLE, M. L.; LEONARDO, R. S.; DWECK, J. Comparative study of biodiesel oxidation stability using Rancimat, PetroOXY, and low P-DSC. **Journal of Thermal Analysis and Calorimetry**, [*s. l.*], v. 116, n. 1, p. 113–118, 2014.

NAIR, Jayashri Narayanan. Study of Biodiesel Blends and Emission Characteristics of Study of Biodiesel Blends and Emission Characteristics of. [*s. l.*], v. 2, n. August 2013, p. 3710–3715, 2015.

PASQUINI, Celio. Near infrared spectroscopy: A mature analytical technique with new perspectives e A review. **Analytica Chimica Acta**, [*s. l.*], v. 1026, 2018.

PELISSON, Leidimara. **Produção de biodiesel por meio de fluidos supercríticos e sua caracterização utilizando Cromatografia Gasosa de Alta Resolução (HRGC)**. 2013. 102 f. [*s. l.*], 2013.

RAMOS, Luiz Pereira *et al.* Biodiesel. **Revista Biotecnologia Ciência & Desenvolvimento**, [*s. l.*], v. 31, p. 28–37, 2003.

REZANIA, Shahabaldin *et al.* Review on transesterification of non-edible sources for biodiesel production with a focus on economic aspects, fuel properties and by-product applications. **Energy Conversion and Management**, [*s. l.*], v. 201, n. July, p. 112155, 2019. Disponível em: https://doi.org/10.1016/j.enconman.2019.112155.Acesso em: 12/12/2021.

SALA, Oswaldo. **Fundamentos da Espectroscopia Raman e no Infravermelho**. 2ºed. [*S. l.*]: Editora Unesp, 2011.

SCHUTTLEFIELD, Jennifer D.; GRASSIAN, Vicki H. ATR-FTIR spectroscopy in the undergraduate chemistry laboratory part I: Fundamentals and examples. **Journal of Chemical Education**, [*s. l.*], v. 85, n. 2, p. 279–281, 2008.

SILVA, Gildo W.B. *et al.* Biodiesel/diesel blends classification with respect to base oil using NIR spectrometry and chemometrics tools. **JAOCS, Journal of the American Oil Chemists' Society**, [*s. l.*], v. 89, n. 7, p. 1165–1171, 2012.

SINGH, Digambar *et al.* A review on feedstocks, production processes, and yield for different generations of biodiesel. **Fuel**, [*s. l.*], v. 262, n. July, p. 116553, 2020. Disponível em: https://doi.org/10.1016/j.fuel.2019.116553. Acesso em: 03/02/2022.

SKROBOT, Vinicius L.; DE SOUSA SANTOS, Caio; BATISTA BRAGA, Jez Willian. Exploratory Analysis of Automotive Diesel Fuel Stability Test Methods by Infrared Spectroscopy and Parallel Factor Analysis. **Energy and Fuels**, [*s. l.*], v. 33, n. 7, p. 6170–6176, 2019.

STUART, Barbara H. **Infrared Spectroscopy: Fundamentals and Applications Analytical Techniques in the Sciences**. [*S. l.: s. n.*], 2004.

SUTTON, Clifton D. 11 - Classification and Regression Trees, Bagging, and Boosting. *In*: RAO, C R; WEGMAN, E J; SOLKA, J L (org.). **Data Mining and Data Visualization**. [*S. l.*]: Elsevier, 2005. (Handbook of Statistics). v. 24, p. 303–329. Disponível em: https://www.sciencedirect.com/science/article/pii/S0169716104240111. Acesso: 09/09/2021.

VALAND, Reema *et al.* A review of Fourier Transform Infrared (FTIR) spectroscopy used in food adulteration and authenticity investigations. **Food Additives and Contaminants - Part A Chemistry, Analysis, Control, Exposure and Risk Assessment**, [*s. l.*], v. 37, n. 1, p. 19–38, 2020. Disponível em: https://doi.org/10.1080/19440049.2019.1675909. Acesso em: 08/07/2022.

VELVARSKÁ, Romana *et al.* Near-infrared spectroscopy for determining the oxidation stability of diesel, biodiesel and their mixtures. **Chemical Papers**, [*s. l.*], v. 73, n. 12, p. 2987–2993, 2019. Disponível em: https://doi.org/10.1007/s11696-019-00852-4. Acesso em: 27/05/2022.

WESTAD, Frank; MARINI, Federico. Validation of chemometric models - A tutorial. **Analytica Chimica Acta**, [*s. l.*], v. 893, p. 14–24, 2015. Disponível em: http://dx.doi.org/10.1016/j.aca.2015.06.056. Acesso em: 16/10/2021.

YAAKOB, Zahira *et al.* A review on the oxidation stability of biodiesel. **Renewable and Sustainable Energy Reviews**, [*s. l.*], v. 35, p. 136–153, 2014. Disponível em: http://dx.doi.org/10.1016/j.rser.2014.03.055. Acesso em: 05/05/2022.

ZHOU, Jian *et al.* Analysis of the oxidative degradation of biodiesel blends using FTIR, UV–Vis, TGA and TD-DES methods. **Fuel**, [*s. l.*], v. 202, p. 23–28, 2017. Disponível em: http://dx.doi.org/10.1016/j.fuel.2017.04.032. Acesso em: 10/11/2022.

ANEXO

Figura 1A – Gráfico de escores para as três primeiras componentes principais (PCs) para dados de NIR

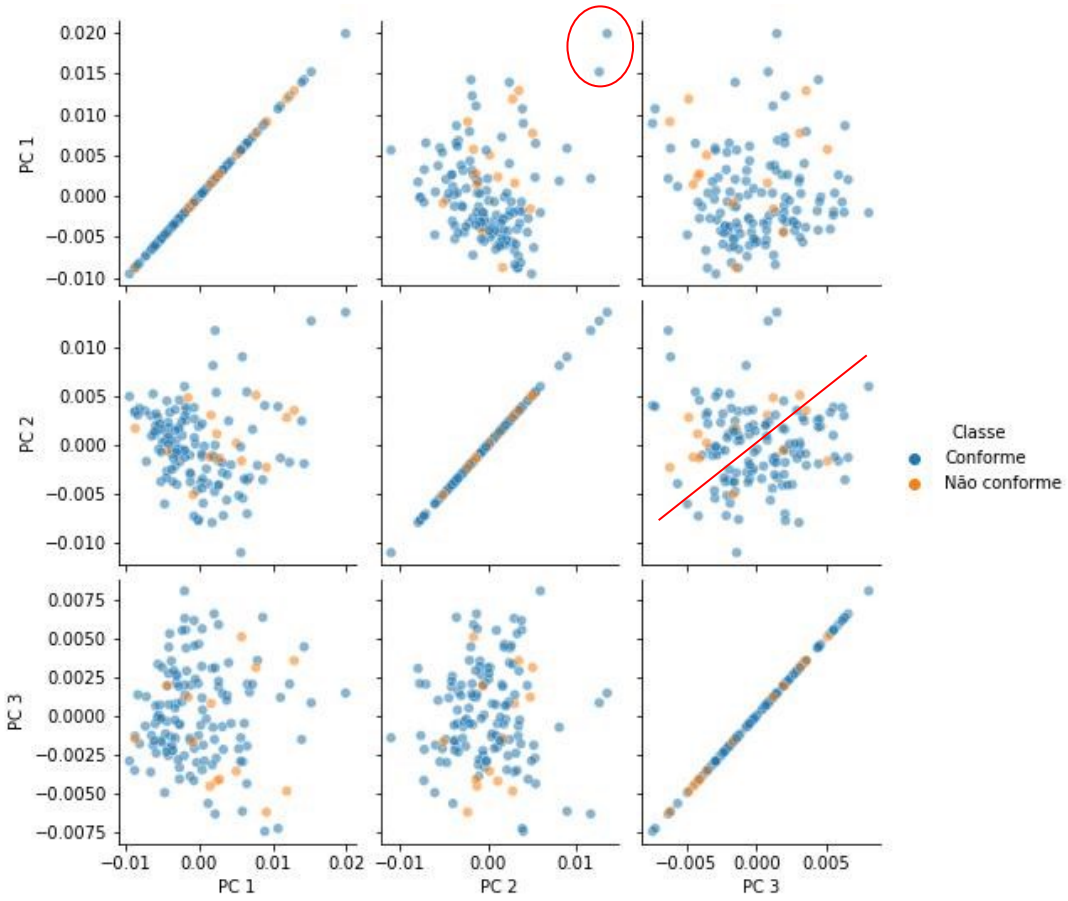


Figura 2A - Gráfico de escores para as três primeiras componentes principais (PCs) para dados de Raman

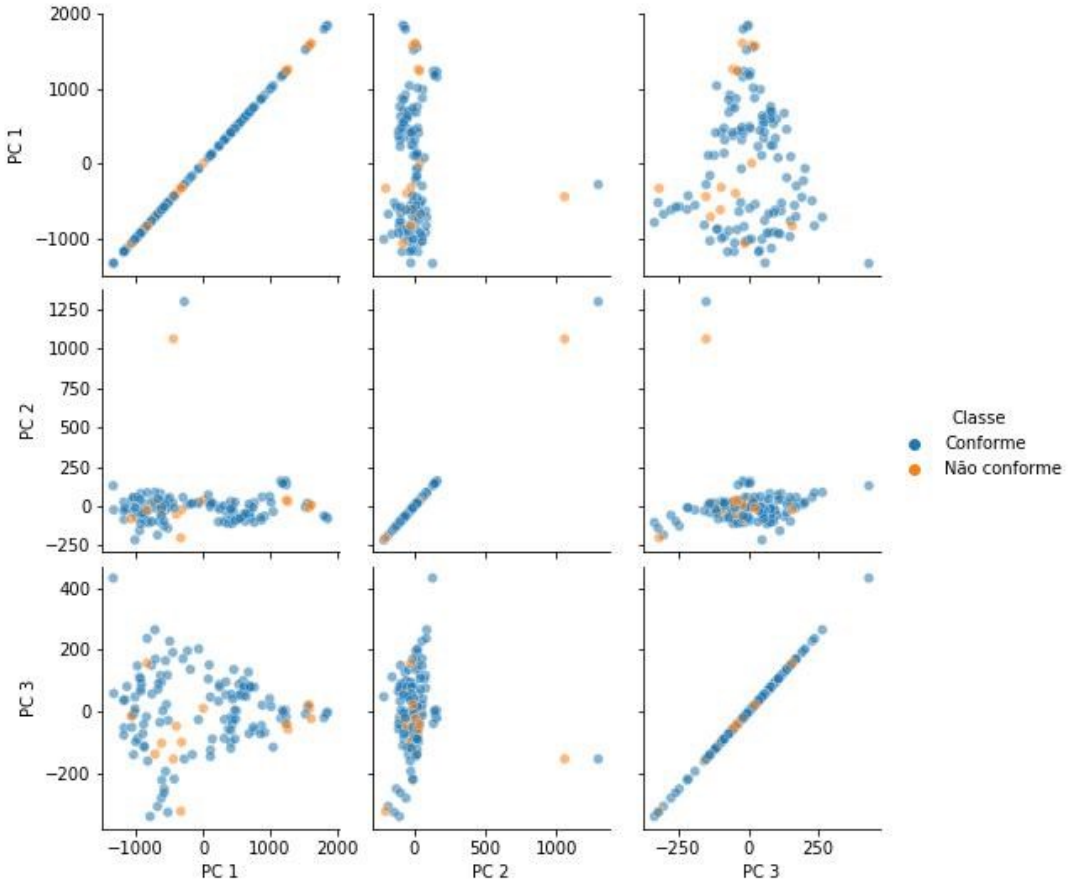


Figura 3A – Coeficientes para o vetor de regressão do LDA sem reamostragem, dados de ATR-FTIR

Figura 4A – Coeficientes para o vetor de regressão do LDA sem reamostragem, dados de NIR

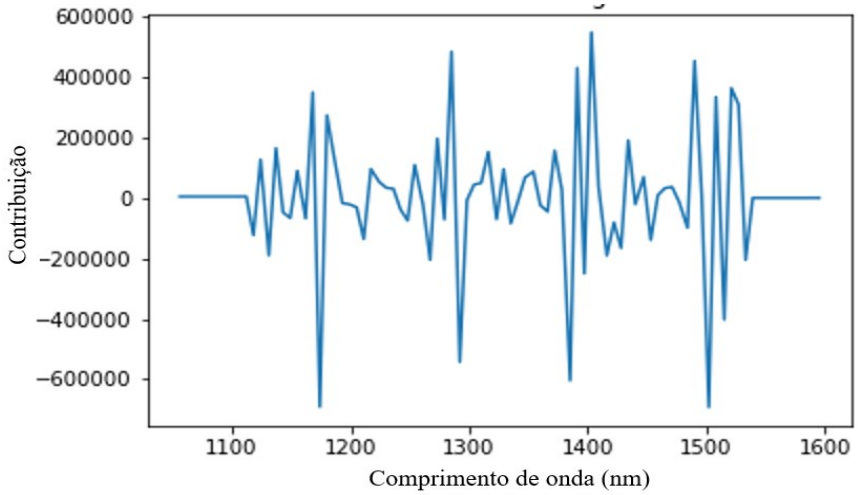


Figura 5A – Coeficientes para o vetor de regressão do LDA sem reamostragem, dados de Raman

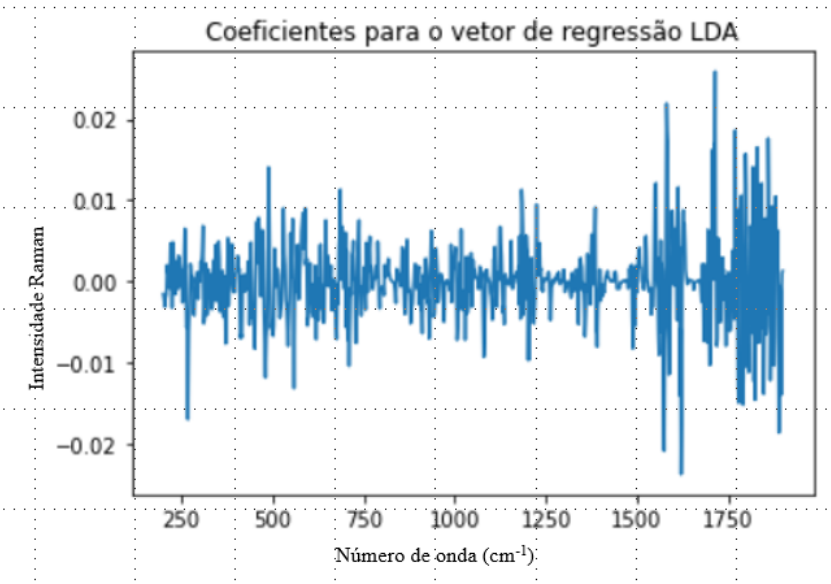


Figura 6A – Importância das variáveis para a floresta aleatória sem reamostragem, dados de ATR-FTIR

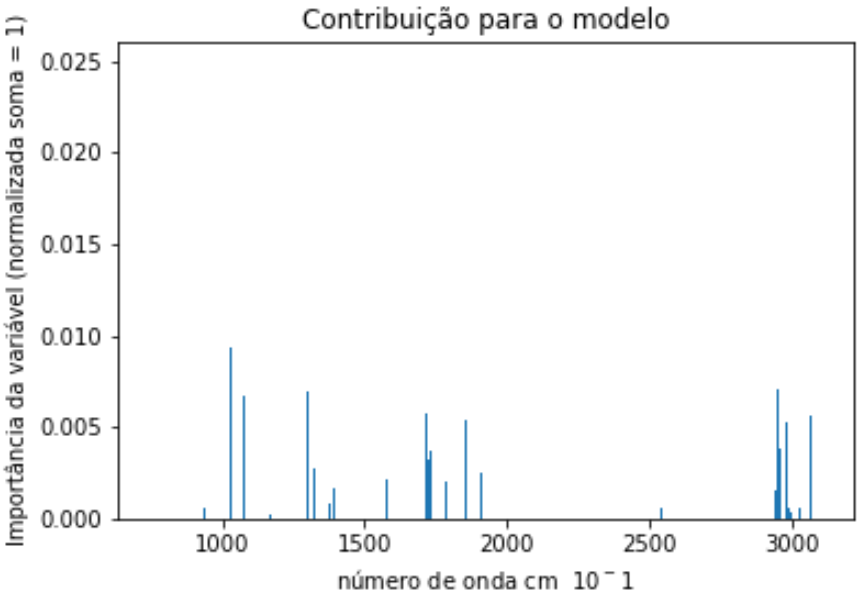


Figura 7A – Importância das variáveis para a floresta aleatória sem reamostragem, dados de NIR

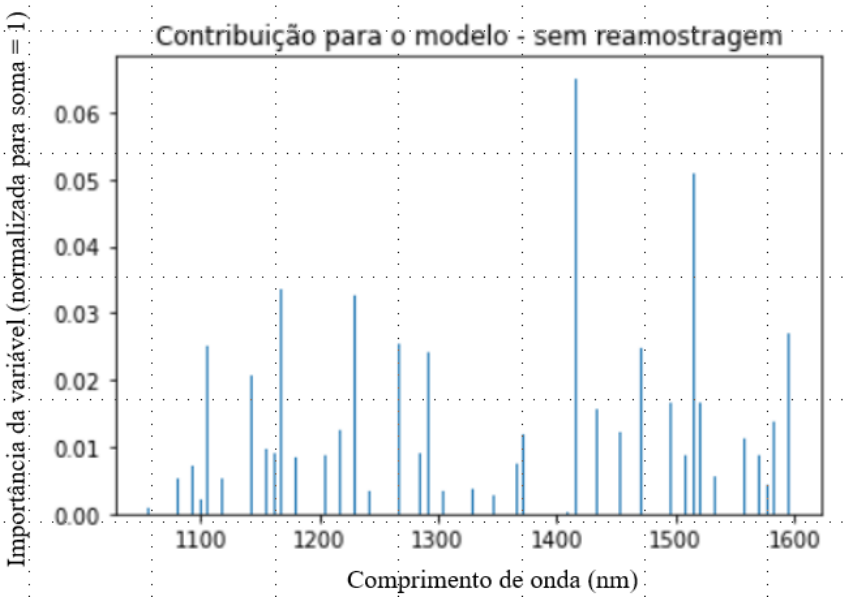
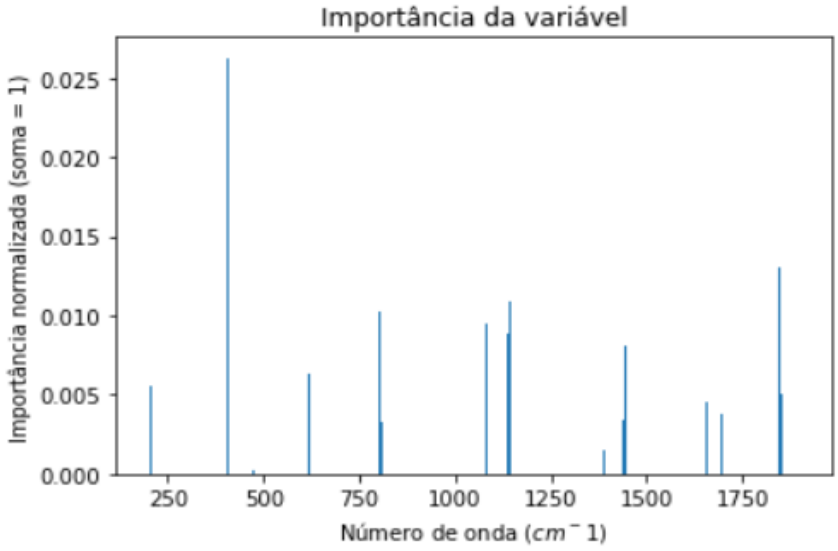


Figura 8A – Importância das variáveis para a floresta aleatória sem reamostragem, dados de Raman sem reamostragem



Quadro 1A – Parâmetros ótimos para a floresta aleatória sem reamostragem dados de ATR-FTIR

Parâmetros	**Valores ótimos**
Número de árvores	70
Profundidade máxima da árvore	2
Variáveis disponíveis a cada construção	$\sqrt[2]{1950}$
Número mínimo de amostras em um nó	2

Quadro 2A – Parâmetros ótimos floresta aleatória sem reamostragem - dados NIR

Parâmetros	**Melhores valores**
Número de árvores	50
Profundidade máxima da árvore	7
Variáveis disponíveis a cada construção	$\sqrt[2]{88}$
Número mínimo de amostras em um nó	2

Quadro 3A – Parâmetros ótimos floresta aleatória sem reamostragem - dados Raman

Parâmetros	**Parâmetros ótimos**
Número de árvores	50
Profundidade máxima da árvore	3
Variáveis disponíveis a cada construção	$\sqrt[2]{942}$
Número mínimo de amostras em um nó	2