UNIVERSIDADE FEDERAL DO ESPÍRITO SANTO

CENTRO DE CIÊNCIAS EXATAS
PROGRAMA DE PÓS-GRADUAÇÃO EM QUÍMICA

Métodos de aprendizagem de máquina em química analítica: Floresta Randômica aplicada na avaliação de petróleo

Machine learning methods in analytical chemistry: Random Forest applied in crude oil evaluation

Betina Pires Oliveira Lovatti

Tese de doutorado em Química

Vitória
2019

Betina Pires Oliveira Lovatti

Métodos de aprendizagem de máquina em química analítica: Floresta Randômica aplicada na avaliação de petróleo

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

Área de Concentração: Química

Linha de Pesquisa: Química do Petróleo e Biocombustíveis.

Orientador: Prof. Dr. Paulo Roberto Filgueiras.

VITÓRIA

2019

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

---

Lovatti, Betina Pires Oliveira, 1986-
L896m Métodos de aprendizagem de máquina em química analítica : Floresta Randômica aplicada na avaliação de petróleo / Betina Pires Oliveira Lovatti. - 2019.
141 f.

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

1. Aprendizagem de máquina. 2. Floresta Randômica. 3. Petróleo. 4. Redução de variáveis. 5. Ressonância magnética nuclear. I. Filgueiras, Paulo Roberto. II. Universidade Federal do Espírito Santo. Centro de Ciências Exatas. III. Título.

CDU: 54

Métodos de aprendizagem de máquina em Química Analítica: Floresta Randômica aplicada na avaliação de petróleo

Betina Pires Oliveira Lovatti

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

Aprovada em 29 de novembro de 2019 por:

Prof. Dr. Paulo Roberto Filgueiras
Universidade Federal do Espírito Santo
Orientador

Prof.ª Dr.ª Mariana Ramos de Almeida
Universidade Federal de Minas Gerais

Prof.ª Dr.ª Emanuele Catarina da Silva Oliveira
Instituto Federal do Espírito Santo

Prof. Dr. Álvaro Cunha Neto
Universidade Federal do Espírito Santo

Prof. Dr. Wanderson Romão
Universidade Federal do Espírito Santo

Universidade Federal do Espírito Santo
Vitória, novembro de 2019

Dedico aos meus pais, irmãos, esposo, sobrinho e ao meu amado filho Francisco.

AGRADECIMENTOS

Desafio tão grande quanto escrever esta Tese, foi utilizar apenas duas páginas para agradecer as pessoas que fizeram parte desta minha trajetória de um pouco mais de 3 anos na Universidade Federal do Espírito Santo.

Inicío os agradecimentos ao meu Deus, detentor de todo conhecimento.

Agradeço a minha família pelo encorajamento e base para prosseguir sem desistir, sou muito feliz por isso. Obrigada pai e mãe que sempre primaram pela minha educação. Meu esposo, Rafael, por entender que tive muitas vezes que priorizar meus estudos, o tempo todo esteve ao meu lado, incondicionalmente. Minha irmã, Betânia, por todas as contribuições e conhecimento com o CorelDraw, sempre me salvando com seu dom artístico. Estêvão, meu irmão, por me passar tranquilidade em momentos de desespero. Ao meu sobrinho, Antônio pela alegria nas horas certas.

Agradeço as amigas, Júlia, Rayza, Karla e Márcia pelo companheirismo. Minha amiga Júlia, de longas datas, por me apresentar a Quimiometria como se fosse algo de simples. Confesso, foi menos pior porque aprendi com você meus primeiros passos na área. Rayza, agradeço os ensinamentos e dicas de traduções. Karla, por se parecer tanto comigo na hora de escrever um artigo, grande parceira. Márcia, parceira de Floresta Randômica, amiga de “quebrar a cabeça” no Matlab.

Não posso deixar de agradecer o carinho dos meus amigos do grupo de Quimiometria, Elisson, Pedro, Gabi, Mariana, Thales, Francine e Layla.

Estendo meus agradecimentos aos demais colegas de parcerias de trabalho. Ao professor Ricardo Kuster pelas oportunidades de trabalhos com produtos naturais.

Agradeço ao meu orientador e coordenador da PPGQUI, Paulo, pelos ensinamentos e paciência aos meus tantos questionamentos. Não posso deixar de agradecer, pela base, pela expansão de conhecimento, por ter concedido liberdade de colocar minhas idéias em prática e que hoje compõem essa Tese. Obrigada por ter confiado em mim quando precisei mudar de cidade, fiz de tudo para manter nossos prazos. Como sua primeira aluna de Doutorado, presenciei sua evolução como profissional a frente de um grupo e hoje admiro a forma como conduz.

Ao Labpetro, pelas análises instrumentais e caracterização das amostras. Em especial, aos colegas da RMN e infravermelho pelo fornecimento dos dados para execução do meu trabalho.

Agradeço também aos funcionários Suzi e Alex, sempre muito solícitos em atender minhas necessidades.

Ao CENPES, PETROBRAS, CNPq, CAPES e FAPES pelo incentivo a pesquisa e apoio financeiro.

Ao PPGQUI-UFES.

Hoje, em especial, agradeço novamente a Deus, por finalizar esse desafio que é o doutorado em Química e poder iniciar um novo desafio que é de ser mãe.

“Os teus olhos viram o meu corpo ainda informe; e no teu livro todas estas coisas foram escritas; as quais iam sendo dia a dia formadas, quando nem ainda uma delas havia.”

Salmos 139:16

Curriculum Vitae

Formação acadêmica

Doutoranda em Química - área Química Analítica, 2019. Universidade Federal do Espírito Santo - UFES, Vitória-ES.

Mestre em química – área Química Orgânica, 2012. Universidade Federal de Viçosa – UFV, Viçosa-MG.

Engenheira Química, 2017. Faculdade Pitágoras, Guarapari-ES.

Licenciada em Química, 2010. Instituto Federal do Espírito Santo – IFES, Vitória-ES.

Atividades acadêmicas

Artigos publicados durante o doutorado

1. **LOVATTI, B. P. O.**; NASCIMENTO, M. H. C.; NETO, A. C.; CASTRO, E. V. R.; FILGUEIRAS, P. R. Use of random forest in the identification of important variables. *Microchem. J*. 2019; **145**: 1129-1134.

2. **LOVATTI, B. P. O.**; SILVA, S. R. C.; PORTELA, N. A.; SAD, C. M. S.; RAINHA, K. P.; ROCHA, J. T. C.; ROMÃO, W.; CASTRO, E. V. R.; FILGUEIRAS, P. R. Identification of petroleum profiles by infrared by spectroscopy and chemometrics. Fuel, **264**, 1-7, 2019.

3. RAINHA, K. P.; ROCHA, J. T. C.; RODRIGUES, R. R. T.; **LOVATTI, B. P. O.**; CASTRO, E. V. R.; FILGUEIRAS, P. R. Determination of API Gravity and Total and Basic Nitrogen Content by Mid- and Near-Infrared Spectroscopy in Crude Oil with Multivariate Regression and Variable Selection Tools. Analytical Letters, **52**, 1-17, -2019.

4. NASCIMENTO, M. H. C.; **OLIVEIRA, B. P.**; RAINHA, K. P.; CASTRO, E. V. R.; SILVA, S. R. C.; FILGUEIRAS, P. R. Determination of flash point and Reid vapor pressure in petroleum from HTGC and DHA associated with chemometrics. Fuel, **234**, 643-649, 2018.

Artigos submetidos durante o doutorado

5. **LOVATTI, B. P. O.**; NASCIMENTO, M. H. C.; RAINHA, K. P.; NETO, A. C.; CASTRO, E. V. R.; FILGUEIRAS, P. R. Different strategies for the use of Random Forest in NMR spectrum. Journal of Chemometrics.

6. LIMA, T. A.; MANHÃES, L. R.; GOMES, A. C.; **LOVATTI, B. P. O.**; SIMAS, N. K.; FILGUEIRAS, P. R.; KUSTER, R. M. Agro residues of Drendanthema x grandiflorum as raw material for a larvicidal product. Phytochemistry Letters.

7. ALVES, D.; GOMES, A. C.; SIMAS, N. K.; HERINGER, O.; ROMAO, W.; **LOVATTI, B**. **P. O.**; SCHERER, R.; FILGUEIRAS, P. R.; KUSTER, R. M. Sugar cane waste products as source of phytotoxic compounds for Agriculture. European Journal of Agronomy.

8. RAINHA, K. P.; **LOVATTI, B. P. O**.; OLIVEIRA, M. L.; BRESSIANI, T. S. C.; SANTANA, I. L.; CASTRO, E. V. R.; PEREIRA, R.; MOURA, P. R. G.; LELIS, M. F. F. Contribuições da experimentação no ensino de quimiometria: Determinação do teor de biodiesel no diesel por espectroscopia no infravermelho. Química Nova.

Artigos para submissão

9. NASCIMENTO, M. H.; **LOVATTI B. P. O.**; FOLLI, G. S.; MORO, M. K.; PAULO, E. H.; CUNHA, P. H. P.; ROMÃO, W.; FILGUEIRAS, P. R. Random Forest to FT-ICR MS variable selection to predict petroleum properties.

Apresentação de trabalhos em Congressos

1. VI ENCAQUI – 29 de outubro a 1 de novembro de 2017 (Participação e apresentação de poster).

2. 19º Encontro Nacional de Química Analítica (ENQA) e 7º Congresso Iberoamericano de Química Analítica – 16 a 19 de sembro de 2018, Caldas Novas, GO (Participação e apresentação de pôster).

3. IV Escola de Inverno de Quimiometria - 20 a 23 de agosto de 2019, Porto Alegre, RS (Participação e apresentação de poster).

4. VII ENCAQUI – 4 a 7 de novembro de 2019 (Trabalho em forma de poster).

Premiação em Congresso

19º Encontro Nacional de Química Analítica (ENQA) e 7º Congresso Iberoamericano de Química Analítica — Trabalho premiado na área de Quimiometria

LISTA DE FIGURAS

Figura 2.1 - Representação esquemática de classificação não supervisionada..... 31

Figura 2.2 - Esquema sobre o princípio básico da PCA..... 33

Figura 2.3 - Representação da organização dos dados na forma de matriz 33

Figura 2.4 - Demonstração gráfica do princípio da PCA 34

Figura 2.5 - Representação esquemática de classificação supervisionada. 39

Figura 2.6 - Sistema de classificação baseado em (a) único classificador e (b) múltiplos classificadores 46

Figura 2.7 - Esquema de uma árvore de decisão..... 48

Figura 2.8 - Fluxograma representativo das etapas seguidas em Floresta Randômica 50

Figura 2.9 - Esquema ilustrativo de reamostragem com reposição..... 53

Figura 2.10 - Número de publicações de RF em 20 anos nas áreas de química. Tópico (palavra-chave): Floresta Randômica..... 53

Figura 3.1 - Construção do modelo de RF..... 74

Figura 3.2 - Gráfico de correlação entre ponto de fluidez, teor de aromáticos (aromáticos e polares) e teor de saturados..... 76

Figura 3.3 - Espectros de RMN de ^{1}H das 105 amostras de petróleos utilizadas nos modelos de classificação. (a) dados originais e (b) após o alinhamento pelo algoritmo icoshift..... 78

Figura 3.4 - Identificação das variáveis preditoras; (a) deslocamento químico no espectro de RMN ^{1}H (b) gráfico da frequência no deslocamento químico..... 80

Figura 3.5 - Espectro de RMN de ^{1}H e peso das variáveis (frequência relativa) por região espectral..... 80

Figura 3.6 - Espectro de RMN de ^{13}C e peso das variáveis (frequência relativa) por região espectral..... 81

Figura 4.1 - Histograma do conjunto amostral em relação ao número de acidez total..... 89

Figura 4.2 - Etapas da construção dos modelos de RF. Usando toda a faixa espectral e com métodos de seleção de variáveis para os dados de RMN..... 90

Figura 4.3 - Espectros de RMN de ^{1}H das 126 amostras de petróleos utilizadas nos modelos de classificação. (a) dados originais e (b) após o alinhamento pelo algoritmo icoshift..... 95
Figura 4.4 - Identificação das variáveis preditoras; (a) deslocamento químico no espectro de RMN de ^{1}H (b) gráfico da frequência no deslocamento químico para modelo de classificação FD e RF..... 97
Figura 4.5 - Espectro de RMN de ^{1}H e peso das variáveis (frequência relativa) por região espectral para modelo de classificação obtido por RF..... 98
Figura 4.6 - Espectro de RMN de ^{1}H e peso das variáveis (frequência relativa) por região espectral para modelo de classificação obtido por FD e RF..... 99
Figura 4.7 - Espectro de RMN de ^{13}C e peso das variáveis (frequência relativa) por região espectral para modelo de classificação obtido por RF..... 100
Figura 4.8 - Espectro de RMN de ^{13}C e peso das variáveis (frequência relativa) por região espectral para modelo de classificação obtido por FD e RF..... 101
Figura A.1 - Gráfico bidimensional (largura e comprimento) das espécies da flor de íris de Fisher..... 125
Figura A.2 - Gráfico bidimensional contendo a primeira partição do espaço 126
Figura A.3 - Construção da árvore de decisão..... 127
Figura A.4 - Gráfico bidimensional contendo a segunda partição do espaço..... 110
Figura B.1 - Histograma de propriedades físico-químicas (gravidade API, teor de saturados, teor de aromáticos e teor de polares) para as 105 amostras de petróleo..... 128
Figura B.2 - Espectro de RMN de ^{13}C das 105 amostras de petróleo usadas nos modelos de classificação. (a) dados originais (b) após alinhamento pelo algoritmo icoshift..... 128

Figura D.1 - Espectros de RMN de ^{1}H das 126 amostras de petróleos utilizadas nos modelos de classificação. (a) dados originais e (b) após o alinhamento pelo algoritmo icoshift..... 131
Figura D.2 - Identificação das variáveis preditoras; (a) deslocamento químico no espectro de RMN de ^{1}H (b) gráfico da frequência no deslocamento químico para modelo de classificação RF.... 132
Figura D.3 - Identificação das variáveis preditoras; (a) deslocamento químico no espectro de RMN de ^{13}C (b) gráfico da frequência no deslocamento químico para modelo de classificação RF.... 132
Figura D.4 - Identificação das variáveis preditoras; (a) deslocamento químico no espectro de RMN de ^{13}C (b) gráfico da frequência no deslocamento químico para modelo de classificação FD e RF..... 133
Figura F.1 - Espectro MIR de amostras de petróleo (a) original e (b) após pré-processamento com SNV..... 111
Figura F.2 - Variância explicada dos modelos de PCA construídos com dados do espectro infravermelho médio (a) e das propriedades físico-químicas das amostras (b)..... 112
Figura F.3 - Gráfico de escores 2D (PC1 e PC2) do PCA usando os espectros MIR das amostras de treinamento..... 113
Figura F.4 - Gráfico de escores 2D (PC1 e PC2). As amostras de teste são destacadas em negrito: (a) amostras t1, t3, t4, t6, t11 e t12; (b) amostras t2, t5, t7, t8, t9 e t10. As elipses representam as cinco amostras mais próximas..... 113
Figura F.5 - Biplot de escores e pesos (PC1 e PC2) para as propriedades físico-químicas de amostras de treinamento e teste (t1 a t12). 116
Figura F.6 - Gráfico de correlação das propriedades físico-químicas de amostras de petróleo..... 117

.

LISTA DE TABELAS

Tabela 3.1 - Resultados dos modelos de classificação para o PFM das 105 amostras de petróleo usando Floresta Randômica..... 79

Tabela 4.1 - Resultados dos modelos de classificação para o NAT das 126 amostras de petróleo usando Floresta Randômica e seleção de variáveis..... 96

Tabela B.1 - Coeficiente de correlação entre propriedades: API, teor de saturados (SAT), teor de aromáticos (ARO) e ponto de fluidez máximo (PFM)..... 129

Tabela B.2 - P-valor no teste de hipótese para significância estatística do cálculo dos coeficientes de correlação. H^0 = não há evidência significativa da correlação entre propriedades..... 129

Tabela C.1 - Assinalamentos estruturais de RMN de ^{1}H..... 130

Tabela C.2 - Assinalamentos estruturais de RMN de ^{13}C..... 130

Tabela D.1 - Tempo gasto na execução dos modelos de classificação..... 131

Tabela E.1 - Propriedades físico-químicas das amostras de treinamento e teste..... 134

Tabela F.1 - Propriedades físico-químicas das amostras de petróleo..... 109

Tabela F.2 - Amostra de teste e respectivas amostras de treinamento por ordem de proximidade..... 114

LISTA DE ABREVIATURAS E SIGLAS

ANP -	Agência Nacional do Petróleo
APPI FT-ICR MS -	Fotoionização à Pressão Atmosférica por Espectrometria de Massas por Ressonância Ciclotrônica de Íons com Transformada de Fourier (*Fourier Transform Ion Cyclotron Resonance Mass Spectrometry*)
ASTM -	Sociedade Americana para Ensaios de Material (*American Society for Testing Materials*)
bagging -	Agregação por *bootstrap* (*bootstrapping Aggregation*)
CART -	Árvore de Classificação e Regressão (*Classification and regression trees*)
CV -	Validação Cruzada (*Cross-Validation*)
ES -	Espírito Santo
FD -	Discriminante de Fisher (*Fisher's Driscriminant*)
FFT -	Transformada Rápida de Fourier (*Fast Fourier Transform*)
GC-MS -	Cromatografia Gasosa Acoplada ao Espectrômetro de Massas (*Gas Chromatography-Mass Spectrometry*)
GC x GC-TOF MS -	Cromatografia Gasosa Bidimensional Abrangente acoplada à Espectrometria de Massas por Tempo de voo (Comprehensive Two-dimensional Gas Chromatography coupled with a Timeof-flight Mass Spectrometer)
HCA -	Análise de Agrupamentos Hierárquicos (*Hierarchical Cluster Analysis*)
ICP-MS -	Espectrometria de Massa com Plasma Indutivamente Acoplado (*Inductively Coupled Plasma Mass Spectrometry*)
ICP-OES -	Espectrometria de Emissão Óptica com Plasma Indutivamente Acoplado (*Inductively Coupled Plasma Atomic Emission Spectrometry*)
IUPAC -	União Internacional de Química Pura e Aplicada (*International Union of Pure and Applied Chemistry*)
k-NN -	k-ésimos Vizinhos mais Próximos (*k-Nearest Neighbor*)
LDA -	Análise de Discriminate Linear (*Linear Discriminant Analysis*)

LIBS -	Espectroscopia de Emissão em Plasma induzido por Laser (*Laser-induced Breakdown Spectrocopy*)
microNIR -	micro- Espectroscopia por Infravermelho Próximo (*micro Near Infrared Spectroscopy*)
MIR -	Espectroscopia na região do Infravermelho Médio (*Mid Infrared Spectroscopy*)
NAT -	Número de Acidez Total
PCA -	Análise de Componentes Principais (*Principal Component Analysis*)
PCA-DA -	Análise Discriminante por Análise dos Componentes Principais (*Principal component analysis discriminant analysis*)
PFM -	Ponto de Fluidez Máximo
PLS -	Mínimos Quadrados Parciais (*Partial Least Squares*)
PLS-DA -	Análise Discriminante por Mínimos Quadrados Parciais (*Partial Least Squares-Discriminant Analysis*)
PPGQUI -	Programa de Pós-Graduação de Química
PVT -	Amostra de Teste de Verificação de Performance
RF -	Floresta Randômica (*Random Forest*)
RMN de ^{1}H -	Ressonância Magnética Nuclear de Hidrogênio (*Nuclear Magnetic Resonance of Hydrogen*)
RMN de ^{13}C -	Ressonância Magnética Nuclear de Carbono (*Nuclear Magnetic Resonance of Carbon*)
RMSECV -	Erro Quadrático Médio de Validação Cruzada (*Root Mean Squared Error Cross Validation*)
SAP -	Saturados, Aromáticos e Polares.
SNV -	Variação Normal Padrão (*Standard Normal Variate*)
SPME-GC-MS -	Microextração em fase Sólida acoplada a Cromatografia Gasosa com Espectrometria de Massas (*Solid-phase Microextraction Followed by Gas Chromatography-Mass Spectrometry*)
SVM -	Máquina de Vetor Suporte (*Support Vector Machine*)
UFES -	Universidade Federal do Espírito Santo

VIP - Variável Importante na Projeção (*Variable Importance in Projection*)

LISTA DE SÍMBOLOS

Arg -	argumento
ºC -	Graus Celsius
^{13}C -	Carbono
CH_2Cl_2 -	diclorometano
Cd -	Cádmio
$CDCl_3$ -	clorofórmio deuterado
d -	Variação intra classe
D -	Variação entre classes
E -	Resíduos
^{19}F -	Flúor
FN -	Falso negativo
FP -	Falso positivo
g -	grama
h -	componentes principais
H -	Hora
^{1}H -	Hidrogênio
Hz -	Hertz (Frequência)
I -	Matriz Identidade
m -	amostras
mg -	miligramas
mg $KOH \cdot g^{-1}$ -	miligramas de hidróxido de potássio por grama.
MHz -	Mega Hertz
min -	minutos
mL -	mililitros
mm -	milímetro
%m/m -	Porcentagem em massa
$mol.L^{-1}$ -	Mol por litro
µs -	microsegundo
n -	variáveis
N -	Número de árvores
^{15}N -	Nitrogênio
Ni -	Níquel

P -	Pesos (*Loadings*)
^{31}P -	Fósforo
p -	autovetores
ppm -	Partes por milhão
s -	segundos
T -	Escores
T -	Tesla
^{T} -	Transposto
TN -	Verdadeiro negativo
TP -	Verdadeiro positivo
µL -	microlitros
V -	Vanádio
Var -	Variância
%v/v -	Porcentagem em volume
X -	Matriz
° -	Graus
√ -	Raiz quadrada
% -	Porcentagem
λ -	autovalores

RESUMO

O desenvolvimento tecnológico impulsionou os laboratórios químicos com instrumentos capazes de extrair cada vez mais informações das amostras. Isto afetou especialmente a área de Química Analítica. A utilização de métodos de estatística multivariada, a qual está inserida dentro de uma crescente área da Química Analítica, denominada Quimiometria, auxilia a exploração de toda a potencialidade destes novos instrumentos. Na vanguarda da Quimiometria está o novo método de aprendizagem de máquina: Floresta Randômica (RF). Este método tem suas aplicações voltadas à modelagem de matrizes complexas tal como o petróleo. A complexidade do petróleo é devida à grande variação na composição de seus constituintes, o que lhe confere propriedades físico-químicas distintas. Essas variações composicionais podem ser observadas por meio de técnicas espectroscópicas tais como: espectroscopia na região do Infravermelho Médio (MIR), Ressonância Magnética Nuclear de Hidrogênio (RMN de ^{1}H) e Carbono (RMN de ^{13}C) que possuem potencialidade de extrair informações a nível molecular do petróleo. Através da aplicação de métodos quimiométricos essas informações químicas podem ser relacionadas a propriedades físico-químicas do petróleo. Assim, o presente trabalho visa a classificação de amostras petróleo utilizando técnicas espectroscópicas associada a métodos de aprendizagem de máquina, assim como, explorar a potencialidade da técnica de RF quando combinada a métodos de seleção de variáveis, além de identificar nesse algoritmo variáveis que mais contribuem para a classificação das amostras de petróleo. Os resultados mostraram que a RF foi capaz de discriminar amostras de petróleo quanto ao Ponto de Fluidez Máximo (PFM) a partir de dados de RMN de ^{1}H e ^{13}C. Ademais, foi possível identificar as variáveis de maior contribuição na modelagem, no qual foi observado um equilíbrio entre compostos aromáticos e saturados. Numa segunda aplicação, RF foi eficiente em discriminar espectros de RMN de ^{1}H e ^{13}C em relação ao Número de Acidez Total (NAT) de petróleo, especialmente, quando associada aos métodos: Análise de Componentes Principais (PCA) e Discriminante de Fisher (FD). A identificação das variáveis mais importantes para discriminação mostrou uma contribuição sutilmente maior da região de aromáticos. Na terceira aplicação, os métodos de reconhecimento de padrões: PCA e k-Vizinhos mais Próximos foram eficientes para identificar perfis de petróleo a partir de dados de MIR, possibilitando obter informações sobre similaridade química de óleos sem a necessidade de uma caracterização completa do óleo.

Palavras-chave: Aprendizagem de máquina, Floresta Randômica, Petróleo, Redução de variáveis; RMN.

ABSTRACT

Technological development has driven chemical laboratories with instruments capable of extracting more information from samples. This has especially affected the area of Analytical Chemistry. The use of multivariate statistical methods, which is part of a growing area of Analytical Chemistry, called Chemometrics, helps to explore the full potential of these new instruments. At the forefront of Chemometrics is the new machine learning method: Random Forest (RF). This method has its applications aimed at modeling complex matrices such as petroleum. The complexity of oil is due to the wide variation in the composition of its constituents, which gives it distinct physicochemical properties. These compositional variations can be observed by spectroscopic techniques such as Mid Infrared (MIR) spectroscopy, Hydrogen Nuclear Magnetic Resonance (^{1}H NMR) and Carbon (^{13}C NMR) that have the potential to extract information at the molecular level of petroleum. Through the application of chemometric methods, this chemical information can be related to the physicochemical properties of petroleum. Thus, the present work aims to classify petroleum samples using spectroscopic techniques associated with machine learning methods, as well as, to explore the potentiality of the RF when combined with variable selection methods, and to identify in this algorithm variables that most contribute for the classification of oil samples. The results showed that RF was able to discriminate petroleum samples according to the Maximum Pour Point (PFM) from ^{1}H and ^{13}C NMR data. Besides, was possible to identify the variables that most contributed to the modeling, in which a balance between aromatic and saturated compounds was observed. In a second application, RF was efficient in discriminating ^{1}H and ^{13}C NMR spectra in relation to the total acidy number (TAN) of oil, especially when associated with the Principal Component Analysis (PCA) and Fisher's Discriminant (FD). The identification of the most important variables for discrimination showed a subtly greater contribution from the aromatic region. In the third application, the pattern recognition methods: PCA and k-Nearest Neighbors were efficient to identify oil profiles from MIR data. This process provides information on the chemical similarity of oils without the need for complete oil characterization.

Keyword: Machine learning method, Random Forest; Crude Oil, Reduction of variables, NMR.

SUMÁRIO

1. INTRODUÇÃO..... 23
1.1 Objetivos..... 25
1.2 Estrutura da Tese..... 25
1.3 Referências Bibliográficas..... 27
2. MÉTODOS DE APRENDIZAGEM DE MÁQUINA EM QUÍMICA ANALÍTICA QUALITATIVA..... 30
2.1 Análise de Componentes Principais – PCA..... 31
2.2 Análise de Agrupamentos Hierárquicos – HCA..... 37
2.3 k-Vizinhos mais Próximos – k-NN..... 38
2.4 Análise de Discriminate Linear – LDA..... 42
2.5 Floresta Randômica – RF..... 47
2.6 Parâmetros de avaliação qualitativos..... 56
2.7 Ressonância Magnética Nuclear e petróleo .....
2.8 Referências Bibliográficas..... 58
3. APLICAÇÃO 1: USO DE FLORESTA RANDÔMICA NA IDENTIFICAÇÃO DE VARIÁVEIS IMPORTANTES..... 68
3.1 Introdução..... 68
3.2 Objetivo..... 70
3.3 Metodologia..... 70
3.3.1 Aquisição de amostras e espectros de RMN 71
3.3.1.1 Aquisição dos espectros do RMN de ^{13}C..... 71
3.3.1.2 Aquisição dos espectros do RMN de ^{1}H..... 71
3.3.2 Medidas do Ponto de Fluidez Máximo..... 72
3.3.3 Quimiometria..... 72
3.3.3.1 Pré-processamento e processamento de dados..... 72
3.3.3.2. Modelo de classificação de Floresta Randômica..... 73
3.3.3.3 Seleção de variáveis..... 75
3.4 Resultados e discussão..... 75
3.5 Conclusões..... 82
3.6 Referências Bibliográficas..... 83

4. APLICAÇÃO 2: DIFERENTES ESTRATÉGIAS PARA O USO DE FLORESTA RANDÔMICA EM DADOS DE RESSONÂNCIA MAGNÉTICA NUCLEAR..... 87
4.1 Introdução..... 87
4.2 Objetivo..... 88
4.3 Metodologia..... 88
4.3.1 Aquisição dos espectros do RMN de ^{13}C..... 90
4.3.2 Aquisição dos espectros do RMN de ^{1}H..... 90
4.3.3 Medidas do número de acidez total 91
4.3.4 Quimiometria..... 91
4.3.4.1 Pré-processamento e processamento de dados..... 91
4.3.4.2 Seleção de variáveis por Análise de Componentes Principais.... 91
4.3.4.3 Seleção de variáveis por Discriminante de Fisher..... 92
4.3.4.4 Modelo de classificação de Floresta Randômica..... 93
4.3.4.5 Identificação de variáveis preditoras..... 93
4.4 Resultados e discussão..... 94
4.5 Conclusões..... 101
4.6 Referências Bibliográficas..... 102
5. CONCLUSÃO GERAL .....
ANEXOS…….…….. 125
Anexo A…….. 125
Anexo B…….. 128
Anexo C…….. 130
Anexo D…….. 131
Anexo E…….. 134
Anexo F……...

1. INTRODUÇÃO

Aprendizado de Máquina (*Machine Learning*, do inglês) é um método capaz de realizar tratamento automatizado de dados analíticos.[1] Esse método revolucionou a área de Química Analítica por proporcionar significado de interesse químico ao grande volume de dados obtidos com a informatização das técnicas instrumentais.[2,3] Neste contexto, a análise de dados para tal nível de complexidade por intervenção da estatística univariada apresentou limitações, dado que, somente uma variável é considerada nas amostras em questão. Nesse caso, cada variável (analito de interesse) é analisada de forma isolada e a presença de interferentes compromete o estudo. Em matrizes complexas, que dependem de muitas variáveis, a aplicação da estatística univariada não é eficiente, visto que não basta conhecer as informações de forma isolada, mas é necessário ter conhecimento da totalidade das informações fornecidas pelo conjunto de variáveis. Em contrapartida, a estatística multivariada considera a correlação entre várias variáveis exploradas simultaneamente, permitindo a aquisição de uma quantidade significativamente maior da informação de interesse. Esses métodos multivariados são vantajosos aos métodos tradicionais univariados porque possibilitam análises de identificação do analito de interesse mesmo em presença de interferentes.[4-6] Por consequência, para a utilização dos métodos multivariados não há necessidade que as amostras sejam padronizadas em laboratório estendendo a aplicação em amostras reais. Essa possibilidade de obter conhecimento químico sobre amostras reais com alto grau de complexidade e poder interpretar de forma clara os dados marca o início da Química Analítica Moderna.

O petróleo é uma matriz complexa em função da variedade de compostos químicos que o constituem.[7-9] A depender da origem, os petróleos podem apresentar alterações significativas na composição que influenciam diretamente as propriedades físico-químicas do óleo.[10] Uma alternativa para a caracterização das propriedades físico-químicas do óleo cru envolve a associação de técnicas espectroscópicas com métodos de aprendizagem de máquina. As técnicas espectroscópicas permitem a aquisição de grande quantidade de informação a nível molecular. Portanto, ferramentas quimiométricas são de fundamental importância para extrair a informação de interesse e encontrar correlações entre as múltiplas propriedades físico-químicas em amostras de petróleo.[11-17] Nesse sentido, a Análise de Componentes Principais (PCA) é destacada como ferramenta mais aplicada para correlacionar diferentes

óleos por similaridade química.[18] A PCA a depender da natureza dos dados pode apresentar limitacões, algumas propriedades físico-químicas de natureza não linear necessitam de métodos de aprendizagem capazes de solucionar essa problemática.[19]

Nesse sentido, na vanguarda da Quimiometria, o método de aprendizagem Floresta Randômica (RF) vem sendo aplicado na classificação de matrizes com alto nível de complexidade incluindo as que possuem características de não linearidade. Dentre as áreas de maior aplicação de RF estão: análise genética, metabolômica, análise de imagens médicas, controle de qualidade de alimentos e mais recente petróleo. Esse método se baseia em *ensembles*, nos quais vários subconjuntos de amostras e variáveis dão origem a múltiplos classificadores (em forma de árvores de decisão). As respostas desses classificadores são combinadas resultando em um único modelo final de classificação.[19,21,22]

Durante a construção das árvores em RF, parâmetros internos são ajustados para cada tipo específico de dados, dentre eles, quantidade de árvores necessárias que melhor descrevem o modelo. Alguns trabalhos apontam que para grandes conjuntos de dados esse número pode ultrapassar 1000 árvores devido a quantidade de informação em estudo.[19,23-25] Entretanto, esse grande aporte de informação pode gerar elevado custo computacional e demandar maior tempo.[26-28] Nesse sentido, métodos que selecionem variáveis de interesse em um grande conjunto amostral, anteriormente a aplicação de RF podem fornecer modelos mais parcimoniosos e rápidos, além de gerar resultados mais promissores. Na literatura, são escassos estudos relacionados a combinação de técnicas de seleção de variáveis com RF, detalhamentos internos do algoritmo RF sobre a identificação das variáveis que mais contribuem na discriminação das amostras e são utilizadas durante a construção dos modelos de classificação.

1.1 Objetivos

- Desenvolver metodologias para problemas qualitativos em amostras de petróleo utilizando técnicas espectroscópicas associada a métodos quimiométricos de classificação.
- Desenvolver metodologia para identificação das variáveis utilizadas durante o processo de construção das árvores de classificação no algoritmo de Floresta Randômica.
- Desenvolver metodologias de seleção de variáveis associada a técnica de Floresta Randômica.
- Desenvolver metodologia para identificar perfis de petróleo utilizando infravermelho médio e k-vizinhos mais próximos.

1.2 Estrutura da Tese

Nesta tese são desenvolvidas metodologias com métodos de aprendizagem qualitativos com ênfase na técnica de Floresta Randômica para a classificação de amostras de petróleo. Essa tese também reporta metodologia para distinção variáveis de maior contribuição na discriminação das amostras de petróleo dentro do algoritmo de Floresta randômica, assim como estratégias metodológicas de técnicas de seleção de variáveis associadas à Floresta Randômica.

No capítulo 2 é descrito o estado da arte dos métodos mais comuns de aprendizagem de máquina aplicados em Química Analítica para a classificação de dados, assim como, a fundamentação teórica da técnica de Floresta Randômica e suas aplicações.

No capítulo 3 é apresentado um algoritmo interno inserido no método de Floresta Randômica para a identificação das variáveis com maior quantidade de informação para a discriminação de amostras de petróleo quanto ao ponto de fluidez máximo. Nesta aplicação foram classificados petróleos com ponto de fluidez máximo maiores e menores que -9 ºC a partir de dados de Ressonância Magnética Nuclear de Hidrogênio (RMN de ^{1}H) e Carbono (RMN de ^{13}C).

No capítulo 4 são apresentadas estratégias para selecionar variáveis anteriormente a aplicação da técnica de Floresta Randômica para classificar amostras de petróleo em relação ao número de acidez total. As técnicas de

análise de componentes principais e Discriminante de Fisher foram utilizadas para a seleção das variáveis de maior relevância nos dados originais obtidos por RMN de ^{1}H e ^{13}C, em seguida os dados foram classificados pelo método de Floresta Randômica.

No capítulo 5 são apresentadas as conclusões gerais da Tese.

1.3 Referências Bibliográficas

1. KOWALSKI, P. C. J.; ISNEHOUR, T. L.; REILLEY, C. N. Investigation of combined patterns from diverse analytical data using computerized learning machines. *Anal. Chem*., **41**, 1949-1953, 1969.

2. FERREIRA, M.; ANTUNES, A. M.; MELGO, M. S.; VOLPE, P. L. O. Quimiometria I: calibração multivariada, um tutorial. *Química Nova,* **22,** 724-731, 1999.

3. VALDERRAMA, P.; BRAGA. J. W. B.; POPPI, R. J. Estado da arte de figuras de mérito em calibração multivariada. *Quimica Nova*, **32**, 1278-1287, 2009.

4. KOWALSKI, B. R.; SEASHOLTZ, M. B. Recent developments in multivariate calibration. *Journal Chemometrics,* **5**, 129, 1991.

5. OLIVIERI, A. C.; FABER, N. M.; FERRÉ, J.; BOQUÉ R.; KALIVAS, J. H.; MARK, H. Uncertainty estimation and figures of merit for multivariate calibration. *Pure Appl. Chem.*, **78**, 633–661, 2006.

6. SENA, M.; POPPI R. J.; FRIGHETTO, R. T. S.; VALARINI, P. J. Avaliação do uso de métodos quimiométricos em análise de solos. *Quimica Nova*, **23**, 547-556, 2000.

7. SPEIGHT, J. G. Handbook of Petroleum Product analysis. 3ª Ed. New York: Wiley-Interscience, 2002. 10 p.

8. ANP - Agência Nacional do Petróleo, Gás Natural e Biocombustíveis - Petróleo e derivados. Disponível em: www.anp.gov.br. Acessado em 03 de outubro de 2019.

9. RIAZI, M. R. Characterization and properties of petroleum fractions. 1ª Ed. Philadelphia: ASTM Stock Number, MNL50, 2005. 20 p.

10. CURIALE, J. A.; FROLOV, E. B. Occurrence and origin of olefins in crude oils. A critical review. *Org. Geochem*. **1**, 397-408, 1998.

11. KHANMOHAMMADI, M.; GARMARUDI, A. B.; GUARDIA, M. Characterization of petroleum-based products by infrared spectroscopy and chemometrics. *TrAC Trends Anal. Chem.*, **35**, 135-149, 2012.

12. PASQUINI, C.; BUENO, A. F. Characterization of petroleum using near-infrared specstrocopy: quantitative modeling for the true bolling ponit curve and specific gravity. *Fuel*, **86**, 1927-34, 2007.

13. FALLA, F. S; LARINI, C.; LE ROUX, G. A. C.; QUINA, F. H.; MORO, L. F. L. NASCIMENTO, C. A. O. Characterization of crude petroleum by NIR. *J. Pet. Sci. Eng*., **51**,127-137, 2006.

14. KALLEVIK, H.; KVALHEIM, O. M.; SJÖBLOM, J. Quantitative determination of saphaltenes and resins in solution by means of near-infrared spectroscopy. Correlations to emulstion stability. *J Colloid Interface Sci.*, **225**, 494-504, 2000.

15. KUPTSOV, A. K.; ARBUZOVA, T. V. A study of heavy oil fractions by fourier-transform near-infrared raman spectroscopy. *Petroleum Chemistry,* **51**, 203-211, 2011.

16. HANNISDAL, A.; HEMMINGSEN, P. V.; SJÖBLOM, J. Group-type analysis of heavy crude oils using vibration spectroscopy in combinations with multivariate analysis. *Ind. Eng. Chem. Res.*, **44**,1349-1357, 2005.

17. BRERETON, R. G. Chemometrics: Data Analysis for the laboratory and Chemical Plant. Chichester: John Wiley & Sons Ltda, 2003. 5 p.

18. SAD, C. M. S.; LACERDA, J. R. V.; FILGUEIRAS, P. R.; RIGONI, V. S.; BASSANE, J. F. P.; CASTRO, E. V. R.; PEREIRA, K. S.; SANTOS, M. F. P. Limitations of the pour point measurement and the influence of the oil composition on its detection using principal component analysis. *Energy Fuels*. **28**, 1686–1691, 2014.

19. LOVATTI, B. P. O.; NASCIMENTO, M. H. C.; NETO, A. C.; CASTRO, E. V. R.; FILGUEIRAS, P. R. Use of random forest in the identification of important variables. *Microchem. J*. **145**, 1129-1134, 2019.

20. LIU W.; LIU, C.; HU, X.; YANG, J.; ZHENG, L. Application of terahertz spectroscopy imaging for discrimination of transgenic rice seeds with chemometrics. *Food Chem.* **210**, 415-421, 2016.

21. LEE, S.; CHOI, H.; CHA, K.; CHUNG, H. Random forest as a potential multivariate method for near-infrared (NIR) spectroscopic analysis of complex mixture samples: Gasoline and naphtha. *Microchem. J.* **110**, 739-748, 2013.

22. CANIZO B. V.; ESCUDERO, L. B.; PÉREZ, M. B.; PELLERANO, R. G. Intra-regional classification of grape seeds produced in Mendoza province (Argentina) by multi-elemental analysis and chemometrics tools. *Food Chem.* **242**, 272-278, 2018.

23. VIGNEAUA, E.; COURCOUX, P.; SYMONEAUX, R.; GUÉRIN, L.; VILLIÈRE, A. Random forests: A machine learning methodology to highlight the volatile organic compounds involved in olfactory perception, *Food Quality and Pref*., **68**, 135–145, 2018.

24. LI, Y.; ZHANG, J. Y.; WANG, Y. Z. FT-MIR and NIR spectral data fusion: a synergetic strategy for the geographical traceability of Panax notoginseng, *Anal. Bioanal. Chem*. **410**, 91–103, 2018.

25. KUCHERYASVSKLY, S. Analysis of NIR using decisions trees and their ensembles. *J. of Anal. and Test*., **2**, 274-289, 2018.

26. MASILI, A.; PULIGHEDDU, S.; SASSU, L.; SCANO, P.; LAI, A. Prediction of physical-chemical properties of crude oils by ^{1}H NMR analysis of neat samples and chemometrics. *J. Chemometr*., **50,** 729-738, 2012.

27. FILGUEIRAS, P. R.; TERRA, L.; CASTRO E. V. R.; OLIVEIRA, L. M. S. L.; DIAS, J. C. M.; POPPI, R. J. Prediction of the distillation temperatures of crude oils using ^{1}H NMR and support vector regression with estimated confidence intervals. *Talanta*, **142**, 197-205, 2015.

28. FILGUEIRAS P. R.; PORTELA, N. A.; SILVA, S. R. C.; CASTRO E. V. R.; OLIVEIRA, L. M. S. L.; DIAS, J. C. M.; NETO, A. C.; ROMÃO, W.; POPPI, R. J. Determination of saturates, aromatics, and polars in crude oil by ^{13}C NMR and support vector regression with variable selection by genetic algorithm. *Energy Fuels*, **30**, 1972-1978, 2016.

2. MÉTODOS DE APRENDIZAGEM DE MÁQUINA EM QUÍMICA ANALÍTICA QUALITATIVA

O uso de recursos tecnológicos para analisar dados químicos cresceu consideravelmente a partir da década de 70, em parte devido à disponibilidade e aos recentes avanços de ferramentas computacionais capazes de suportar cálculos laboriosos. Em contrapartida, a obtenção de dados, principalmente na área da química analítica, atingiu um alto nível com a conexão instrumental aos computadores gerando grandes volumes de informações com diferentes tipos de complexidade.[1] Assim, tornou-se imprescindível o uso de novas ferramentas para tratamento e extração de informações relevantes destes conjuntos de dados. Nesse contexto, surgiu a Quimiometria, uma área da Química especificamente destinada a analisar dados químicos de natureza multivariada, utilizando recursos matemáticos, estatísticos e químicos viabilizando a obtenção de mais informações quando comparado a procedimentos univariados.[2]

A introdução histórica se inicia (1971) pelo nome “Quimiometria” (*Chemometrics*, do inglês), criado a partir da combinação de duas palavras suecas *Kemo* (Química) e *metri* (medida), pelo professor suéco Svante Wold. No ano seguinte, Wold forma o primeiro grupo de pesquisa em Quimiometria. Semelhante a definição atual, o grupo conceitua, em 1974, esta ferramenta multivariada como “A arte de extrair informação química relevante de dados produzidos em experimentos químicos é chamada de quimiometria em analogia a biometria e econometria”.[3] Entre 1971 e 1975, foram publicados os primeiros artigos que tiveram impacto na história da Quimiometria: Lawton e Sylvestre (1971), Kowalski e Bender (1972), porém, foi apenas em 1975 que o professor Bruce Kowalski publica o primeiro artigo contendo no título o termo quimiometria e uma abordagem conceitual do tema, além disso, define essa ferramenta como uma nova área da Química.[3,4] Finalmente, o reconhecimento dessa nova área da química se concretiza no meio científico com o lançamento de jornais específicos: o *Chemometrics and Intelligent Laboratory Systems* (1986 – Elsevier) e o *Journal of Chemometrics* (1987 – Wiley).[3] Nessa mesma década, professor Kowalski ministra o primeiro curso intensivo das principais vertentes da quimiometria no Instituto de Química da Unicamp, a pedido do professor Roy Bruns, iniciante na área e pioneiro da quimiometria no Brasil.[5]

Dentre as principais vertentes da quimiometria estão: planejamento e otimização de experimentos, calibração multivariada e classificação, destaca-se a análise multivariada de classificação também conhecida como reconhecimento de padrões ou análise exploratória de dados.[6] Essa área é destinada ao tratamento de dados de natureza qualitativa e tem por finalidade discriminar e encontrar padrões em amostras de acordo com suas similaridades. Estes padrões relacionam a identidade de uma amostra com suas características químicas.[7]

Em reconhecimento de padrões, a informação química contida nas amostras é utilizada para verificar a similaridade que pode haver entre elas. O resultado desse processo revela a presença ou ausência de agrupamentos entre as amostras e pode ser representado graficamente. Normalmente, são obtidos gráficos bidimensionais que facilitam a interpretação dos resultados. O reconhecimento de padrões pode ocorrer em dois modos: o supervisionado e o não supervisionado. No modo não supervisionado (Figura 2.1) não há suposição inicial sobre a estrutura dos dados, formando-se agrupamentos naturais com base na informação contida nos dados experimentais em questão.[8] Nesse modo, o principal e mais antigo algoritmo desenvolvido para elaborar gráficos que representem a maior quantidade possível das informações contidas em um conjunto de dados é a Análise de Componentes Principais (PCA – *Principal Component Analysis*).[9]

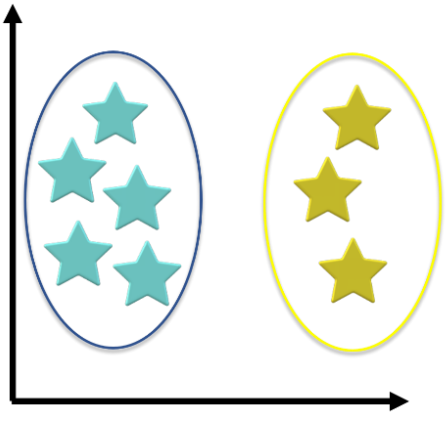

Figura 2.1 Representação esquemática de classificação não supervisionada.

2.1 Análise de Componentes Principais – PCA

A PCA teve sua primeira abordagem teórica em 1901, na Inglaterra com o professor Karl Pearson e seus alunos, dentre eles Fisher, que publicou juntamente

com Mackenzie (1923) o primeiro artigo envolvendo a PCA aplicada na discriminação de diferentes espécies de batatas mediante ao uso de fertilizantes.[10] Essa técnica teve maior impacto com o artigo publicado por Hotteling, em 1933, sobre a teoria envolvendo o redimensionamento de dados que é a base da PCA.[11,12] Na década de 70, com o advento da quimiometria e associação com o método da PCA conduziu à consolidação e à popularização da técnica. O primeiro programa quimiométrico, “ARTHUR”, desenvolvido por Kowalski, em 1975, facilitou a execução de cálculos relacionados a análise exploratória.[5,13] O novo pacote computacional foi importado para o Instituto de Química da Unicamp, em 1978, pelo professor Bruns.[5] Essa fusão impactou a Química Analítica possibilitando resultados mais rápidos e menos laboriosos, sendo essa uma grande vantagem do uso de métodos multivariados em tratamentos de dados em relação aos univariados.

A teoria proposta por Hotteling (1933) sobre o uso da PCA para o redimensionamento teve uma abordagem mais avançada em 1987, no artigo tutorial publicado por Wold et al. Nesse artigo, são definidos conceitos e aplicações da PCA em dados multivariados, pela primeira vez, mostrando que o processo de redimensionamento é obtido ao se projetar os dados multivariados em um subespaço (menor dimensão), consequentemente, ocorre a redução do espaço original do conjunto de dados, sem que a relação entre as amostras seja afetada.[14,15] Em outras palavras, conceitos mais recentes definem que a PCA é aplicada na transformação de grandes matrizes de dados em poucas variáveis, denominadas componentes principais, as quais retém a maior parte da variância (informação). Esta transformação (Figura 2.2) tem por objetivo separar a informação relevante daquela que não é relevante (ruído).[13]

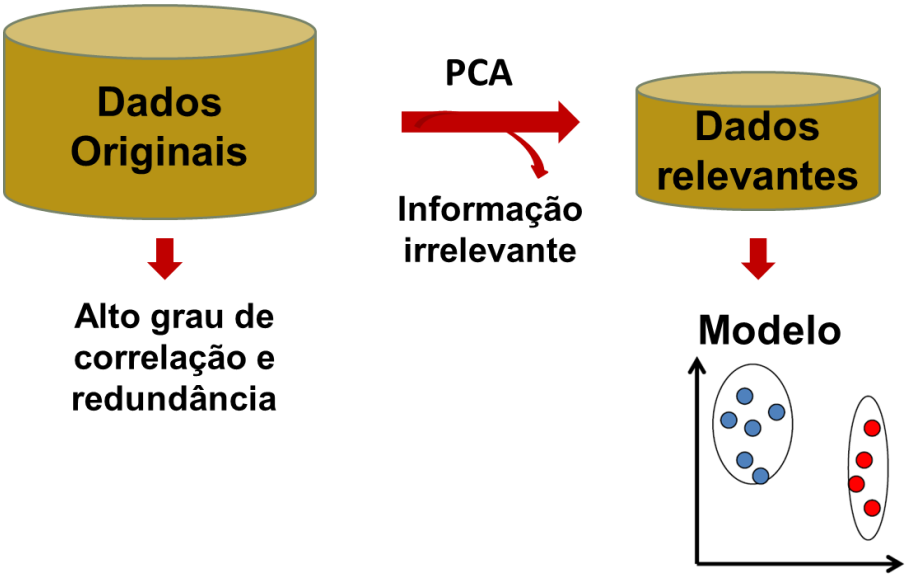

Figura 2.2 Esquema sobre o princípio básico da PCA.

A PCA inicia-se com a organização de dados da matriz "**X**" (Figura 2.3) contendo o conjunto original de variáveis, cujas linhas representam as amostras e as colunas as variáveis (propriedades físico-químicas, número de onda, deslocamento químico, entre outros a depender da instrumentação utilizada).

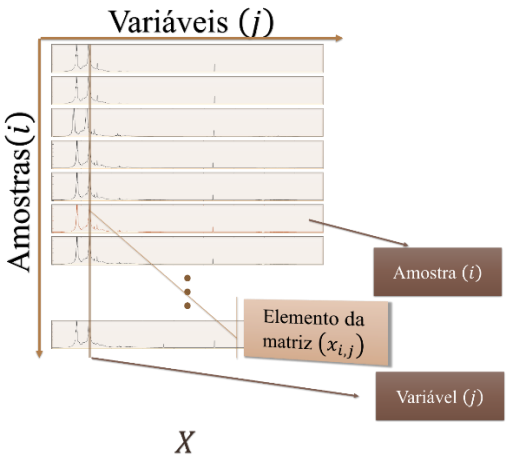

Figura 2.3 Representação da organização dos dados na forma de matriz.

O arranjo de dados da matriz "**X**" pode ser visualizada após sua decomposição em dois modos (Figura 2.4). O primeiro modo contém informações a respeito das amostras (*m*), representado pela matriz de escores (**T**) e no segundo modo contém informações a respeito das variáveis (*n*), representado pela matriz de pesos ou *loadings* (**P**), a informação residual (sem relevância) desse processo é encontrada na matriz de resíduos (**E**) conforme expresso pela Equação (2.1).[9]

$$\mathbf{X}_{(n,m)} = \mathbf{T}_{(n,h)} \cdot \mathbf{P}^{T}_{(m,h)} + \mathbf{E}_{(n,m)} \qquad (2.1)$$

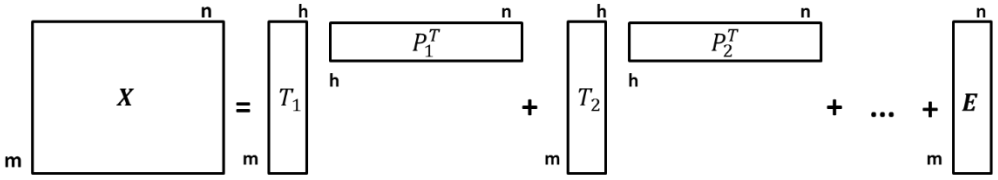

Figura 2.4 Demonstração gráfica do princípio da PCA.

Na PCA ocorre uma transformação dos dados em mesma dimensão, mas, transformado em sistema um ortogonal (independentes). A dimensionalidade dos dados é naturalmente reduzida pela maximização da variância, dado pelas Equações (2.2) e (2.3).[9]

$$\mathbf{P} = \arg\{\max[Var(\mathbf{X})]\} \ldots\ldots\ldots\ldots\ldots \quad (2.2)$$

sujeito a,

$$p_i^T \cdot p_j = \begin{cases} 1, & se \quad i = j \\ 0, & se \quad i \neq j \end{cases} \text{, ou } p^T.p = I \quad (2.3)$$

Na solução desse problema encontra-se uma equação de autovalores (λ) e autovetores (**p**) que são argumentos que maximizam a variância de **X**. Os autovetores, ou pesos são uma combinação linear das variáveis originais, ou seja, possuem informações sobre as variáveis com maior importância em cada autovetor. Estas novas variáveis são chamadas de componentes principais (*h*), Equação (2.1), e são caraterizadas por serem linearmente independentes. Em um modelo PCA espera-se reduzir a dimensão para 2 ou 3 componentes, os quais possuem a maior parte da informação contida no conjunto de dados. Os componentes principais são dispostos em ordem decrescente de variância explicada. A decomposição final da matriz **X** está representada na Equação (2.1).[9]

A redução da dimensionalidade facilita a observação de agrupamentos ou não de amostras por similaridade, ou seja, proximidade entre amostras no espaço. Amostras próximas, no espaço das componentes principais, tendem a ter características semelhantes.[9,13,16]

A possibilidade de visualizar os agrupamentos naturais devido a semelhança química entre as amostras proporcionou classificar conjuntos de dados em diferentes áreas da química. No ramo da química de alimentos, Mildner-szudlarz e Jelen (2008)[17] usaram a PCA como ferramenta no controle de qualidade de azeites comerciais adulterados ou não com óleo de avelã, concluiram que a PCA foi eficiente na classificação das amostras para as diferentes técnicas analíticas usadas para aquisição de informações químicas. O artigo de Lyra et al. (2010)[18] se destaca por classificar dados conhecidos para os químicos e que já possuem uma pré-classificação concedida pela União Internacional de Química Pura e Aplicada (IUPAC - *International Union of Pure and Applied Chemistry*). Nesse trabalho, com enfoque didático, o objetivo principal dos autores foi de agrupar os elementos químicos em classes (metais, ametais, gases nobres e semimetais) e comparar com a classificação definida pela IUPAC tendo como variáveis oito propriedades periódicas dos elementos químicos. Os autores verificaram resultados satisfatórios pelo gráfico bidimensional de escores: com 58% de variância explicada, os elementos químicos foram discriminados em suas respectivas classes químicas. Sabino et al. (2012)[19] aplicaram a PCA em dados de produtos naturais, em que amostras de oito populações de orégano obtidas de diferentes localizações geográficas da Guatemala, foram discriminadas em três quimiotipos.

Essa técnica quimiométrica tem sido empregada em potencial para reconhecimento de padrões em amostras mais complexas como, por exemplo, o petróleo.[20,21] A depender do tipo de óleo, composição química e suas propriedades, alterações de comportamento comprometem atividades de refino, produção e transporte.[22,23] Nesse sentido, a aplicação da PCA tem sido eficiente para obter informações geoquímicas como origem, processos de migração e estudos genéticos, os quais podem influenciar no valor comercial do petróleo.[24] Chiaberge et al. (2013)[25] classificaram amostras de petróleo bruto extraídas de diferentes poços através da PCA associada à fotoionização à pressão atmosférica por espectrometria de massas por ressonância ciclotrônica de íons com transformada de Fourier (APPI FT-ICR MS - *Fourier Transform Ion Cyclotron Resonance Mass Spectrometry*). Os autores concluíram que a caracterização a nível molecular do petróleo combinada a quimiometria possibilitou a discriminação das amostras por origem de poço devido às diferenças nas composições químicas.

Vieira et al. (2016)[24] aplicaram a PCA em dados de densidade API, concentrações de vanádio (V) e níquel (Ni) para determinar origem de amostras de petróleo em diferentes campos da bacia sedimentar do Espírito Santo. Os autores verificaram que a técnica quimiométrica aplicada aos dados de teor de V, Ni e gravidade API foi eficiente na discriminação das amostras em quatro campos, dois de origem marinha e dois de origem lacustre. Observaram também que há uma tendência do Ni se concentrar em petróleos mais pesados, complexados aos asfaltenos e resinas.

Em meio as várias aplicações da PCA em tratamento de dados químicos, Bro e Smilde, em 2014, publicaram o primeiro artigo de revisão sobre a PCA com enfoque quimiométrico. Os autores descrevem aspectos históricos, conceitos, fundamentos teóricos, as diferentes formas de aplicação desta técnica, os recursos que essa ferramenta oferece e como interpretar os resultados quando os dados são químicos.[16]

Outros estudos, não citados no artigo de revisão buscaram desenvolver metodologias mais avançadas combinando quimiometria e dados obtidos de instrumentos portáteis para obterem respostas em menor tempo e tomadas de decisões mais rápidas.[26] Altinpinar et al. (2013)[27] analisaram através da micro-Espectroscopia por Infravermelho Próximo (microNIR – *micro Near Infrared Spectroscopy*) amostras de solos contaminadas com hidrocarbonetos de petróleo. Ao aplicar a PCA verificaram os agrupamentos das amostras de acordo com o tipo de contaminante. Os autores concluíram que a combinação das técnicas pode antecipar resultados e intervenções caso os solos sejam contaminados com petróleo decorrentes de acidentes durante transporte e vazamentos, reduzindo riscos ao meio ambiente.

Correia et al. (2018)[28] usaram microNIR e PCA obtiveram resultados satisfatórios ao analisar tendências em amostras de café brasileiro quanto ao teor de adulterantes. Amostras com maior teor de adulterante foram agrupadas separadamente das amostras com concentrações inferiores.

A versatilidade em aplicações da PCA na solução de problemas envolvendo dados químicos a tornaram a técnica de classificação não supervisionada mais conhecida e consolidada na literatura. Em contrapartida, outros métodos de inferior popularidade, conhecidos como Análise de Agrupamentos Hierárquicos (HCA –

Hierarchical Cluster Analysis) e k-ésimos Vizinhos mais Próximos (k-NN - k-*Nearest Neighbor*) foram também aprimorados pela capacidade de encontrar padrões.

2.2 Análise de Agrupamentos Hierárquicos – HCA

A técnica de HCA teve sua origem na taxonomia numérica em que biólogos estudavam semelhança entre organismos de diferentes espécies.[13] Carolus Linnaeus (1735)[29] foi quem introduziu o primeiro sistema hierárquico sucinto para organizar as espécies de acordo com as suas caraterísticas morfológicas. Nos sistemas hierárquicos, como HCA, os agrupamentos são organizados na forma de dendrogama em uma estrutura hierárquica de dados. Os comprimentos dos ramos representam o grau de similaridade química entre as amostras.[9] Semelhante a PCA, a técnica de HCA teve expansão com o desenvolvimento tecnológico e de algoritmos, na década de 70, possibilitando seu uso em tratamento de dados multivariados.

Em 1981, foi publicado um dos primeiros artigos da aplicação de HCA em análise multivariada de dados químicos. Nesse artigo, Adamson e Bawden (1981)[30] discriminaram substâncias químicas de acordo com o tipo de substituinte e seus efeitos usando dados de atividade biológica. Os autores concluíram que o método de HCA foi eficiente e rápido na discriminação das moléculas de acordo com sua atividade. Na década de 90, a HCA passou a ser associada a dados de espectroscopia para resolver problemas de classificação. Em 1992, Hobert e Meyer usaram o infravermelho associados a HCA para classificar cálculos urinários de diferentes composições químicas. Os autores concluíram que através da técnica instrumental de baixo custo juntamente com a quimiometria foi possível obter resultados de grupos de cálculos renais segundo a constituição química.[31] A partir de então a combinação de técnicas analíticas, principalmente, a espectroscopia e espectrometria com a HCA facilitou a visualização e interpretação de resultados, especialmente, nas áreas de química de alimentos.

Na química de alimentos há maior interesse por parte dos pesquisadores em utilizar essas técnicas para obter resultados relacionados a controle de qualidade. Recentemente, Barbosa et al. (2019)[32] analisaram metais acumulados em tecidos moles de espécies de moluscos bivalves por Espectrometria de Emissão Óptica com Plasma Indutivamente Acoplado (ICP-OES - *Inductively Coupled Plasma Atomic Emission Spectrometry*) e Espectrometria de Massa com Plasma Indutivamente

Acoplado (ICP-MS - *Inductively Coupled Plasma Mass Spectrometry*). No total foram usadas 66 amostras e 12 variáveis. A análise exploratória dos dados foi realizada por HCA e PCA para comparação de resultados entre as duas técnicas, anteriormente à construção dos modelos os dados foram tratados com autoescalamento. Para ambas o resultado mostrou uma tendência de formação de agrupamentos e caracterizações distintas entre as amostras das diferentes espécies de moluscos. Também foi notada uma região de confluência entre as espécies. Os autores concluíram que as duas técnicas foram eficientes na classificação de espécies com base na concentração de metais relacionadas a cada uma delas, também concluíram que esses resultados chamaram a atenção para grupos que apresentavam metais acima da legislação. Lima et al. (2019)[33] verificaram a influência da forma de cozimento (banho-maria, forno e em forno micro-ondas) sobre a composição mineral de folhas de mostarda. A partir dos valores de concentração dos minerais obtidos por ICP-OES das amostras de folhas de mostarda associados às técnicas de HCA e PCA, os autores verificaram que as amostras cruas e cozidas em banho-maria se agruparam separadamente das amostras cozidas em forno ou forno microondas. Como conclusão, os autores apontaram que os minerais são alvo de investigação quanto aos seus efeitos em humanos e que os resultados de discriminação possibilitaram relacionar os minerais presentes na planta ao tipo de cozimento fornecido. Cebi et al. (2019)[34] estudaram a influência da fonte da gelatina (porco ou bovino) usadas em doces tipo goma produzidas nas indústrias. Para discriminar os doces em grupos de acordo com a origem da gelatina foram utilizados dados de espectroscopia de infravermelho próximo e métodos quimiométricos (HCA e PCA). Os dados foram pré-processados usando a primeira derivada para ambas as técnicas. Os resultados mostraram 100% de discriminação dos doces.

2.3 k-Vizinhos mais Próximos – k-NN

Outro classificador com aplicações e objetivos semelhantes aos métodos anteriores (PCA e HCA) é o k-NN. Essa uma das poucas ferramentas multifuncionais presentes na quimiometria, com aplicabilidades no modo não supervionado e mais comumente no modo supervisionado.

O modo supervisionado (Figura 2.5) de reconhecimento de padrões é aquele que a partir de um conjunto de dados original assume-se dois conjuntos, treinamento

e teste, cujas características são conhecidas. Em tratamento de dados químicos essas características são informações de natureza qualitativa, por meio das quais são atribuídas as classes.[8,9]

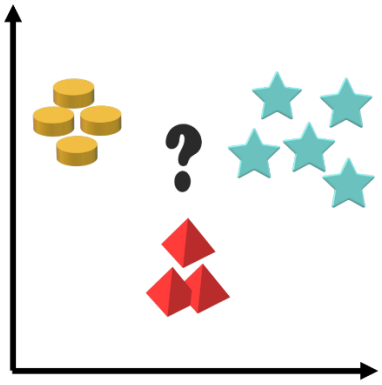

Figura 2.5 Representação esquemática de classificação supervisionada.

A partir do conjunto de treinamento contendo informações qualitativas, os modelos são construídos e possíveis padrões são estabelecidos, (Figura 2.5), representados pelas diferentes formas geométricas. Durante o processo de construção do modelo, cada amostra do conjunto de treinamento é excluída uma única vez e então classificada usando-se para isso as amostras restantes. Para a técnica do k-NN, nessa etapa são calculadas as distâncias entre a amostra excluída das demais amostras constituintes do conjunto de treinamento, quanto mais similares quimicamente forem as amostras, menores serão suas distâncias no espaço. Essas distâncias são organizadas em ordem crescente para facilitar a identificação dos k vizinhos mais próximos. Finalmente, a amostra excluída é então classificada de acordo com a maioria dos “votos” de seus vizinhos mais próximos. Quando a classe designada para amostra excluída coincide com a sua classe verdadeira tem-se êxito na classificação. Esse processo, denominado validação cruzada, é repetido e aplicado para todas as amostras que compõem o conjunto treinamento, na construção final do modelo são obtidos os resultados de êxito e erros de classificação. A vantagem do método supervisionado em relação ao modo não supervisionado é a possibilidade da inserção de amostras desconhecidas no modelo de classificação construído. Como resposta, essas novas amostras são classificadas de acordo com os padrões existentes.[8,9,13,35]

Essas definições do modo supervisionado aplicado ao k-NN foram estabelecidas na década de 70, com as publicações de Patrick e Fischer (1970) e de

Kowalski e Bender (1972), sendo este último artigo voltado para tratamento de dados químicos.[36] Em 1974, Pichler e Perone relataram a principal limitação dessa técnica, em não estabelecer critérios de relevância de informação o que impossibilita solucionar problemas de redução de dimensionalidade dos dados. Nesse caso, as variáveis assumem igual importância.[37] Brereton, em 2003, enfatizou que os dados químicos quando analisados por espectroscopia fornecem muitas variáveis com diferentes níveis de informações, e nesse caso, há duas formas de contornar a problemática, uma delas é selecionar as variáveis, a outra é realizar essa redução da dimensionalidade por meio de outros algoritmos, por exemplo, a PCA.[9] Essa junção de técnicas quimiométricas foi um artifício utilizado por Tsai et al. (2017).[38] Os autores desenvolveram um trabalho de análise exploratória de dados para discriminar diferentes explosivos plásticos. Os dados foram obtidos por Cromatografia Gasosa Bidimensional Abrangente acoplada à Espectrometria de Massas por Tempo de voo (GC x GC-TOF MS - *Comprehensive Two-dimensional Gas Chromatography coupled with a Timeof-flight Mass Spectrometer*) e a eles foram aplicados a PCA para a redução da dimensão dos dados, em seguida, o k-NN para a análise discriminativa. Os resultados mostraram que as duas técnicas quimiométricas combinadas foram úteis nessa classificação. Os autores concluíram que a partir de técnicas quimiométricas simples combinadas foi possível resolver o problema de alta complexidade gerada pelos dados de cromatografia bidimensional.[38]

Slavov et al. (2017)[39] realizaram a classificação de amostras contendo diferentes sorotipagens de Salmonella por meio de dois modos: no primeiro efetuaram o reconhecimento de padrões com todas as variáveis (íons) obtidas por Cromatografia Gasosa Acoplada ao Espectrômetro de Massas (GC-MS - *Gas Chromatography-Mass Spectrometry*) e quimiometria. No segundo modo aplicaram seleção de variáveis nos dados do espectro de massas anteriormente a classificação por k-NN. Os resultados no primeiro modo não foram satisfatórios na classificação. No segundo modo, do total, dezenove íons específicos foram escolhidos para serem usados na classificação de amostras de Salmonella quanto a tipagem. Seis pares de tipagem distintas de Salmonella foram discriminadas em quatro classes. Em seguida, foram testadas amostras desconhecidas de sete amostras externas, dessas seis foram classificadas na tipagem correta. Os autores concluíram que a redundância de informação contida nos íons foi minimizada com a seleção de variáveis, além disso, notaram que os íons

escolhidos estavam relacionados com aminoácidos presentes na membrana celular e que são característicos para cada tipagem de Salmonella, tal fato resultou em bons resultados durante a análise exploratória por k-NN.

Além da limitação da técnica, Pichler e Perone, apontaram como desvantagem que o k-NN apresenta elevado custo computacional por executar muitos cálculos de distancias entre amostras.[37] Porém, com os avanços tecnológicos possibilitou que houvesse redução no custo computacional exigido pelo k-NN, sendo recentemente qualificado como uma técnica simples e de menor custo.[13] Em 2015, Ferreira cita como desvantagem que o k-NN não possui habilidade em identificar amostras com comportamento anômalo, nos conjuntos treinamento e teste. Para contornar essa problemática a PCA pode ser utilizada para detectar amostras atípicas. Em contrapartida, como vantagem, Ferreira, aborda a capacidade da técnica em classificar mesmo em situações em que a quantidade de amostra que compõe cada classe do conjunto de treinamento é desigual.[13]

A maior vantagem desse método está relacionada a sua versatilidade em solucionar problemas de classificação contendo dados de natureza linear como também não linear.[37,40] Essa característica de esclarecer classificações não lineares, até o momento, não havia sido identificada nos métodos anteriormente aqui abordados. No entanto, na literatura, trabalhos envolvendo aplicações do k-NN como método não-linear para tratamento de diferentes matrizes químicas são menos comuns de serem encontrados. Esse fato pode ser justificado em decorrência de métodos não lineares mais sofisticados e com maior poder de discriminação estarem sendo desenvolvidos, dentre eles, destaca-se a Floresta Randômica que será abordada adiante.

Como método não linear, Rodríguez et al. (2019)[40] publicaram um artigo comparando a eficiência do k-NN com outros métodos não lineares mais robustos, desenvolvidos recentemente, para discriminar variedades de feijão. O k-NN foi a técnica que mostrou menor potencialidade durante a classificação. Porém, como método linear, o k-NN tem mostrado capacidade em classificar dados químicos que compõem diversas matrizes em diferentes áreas químicas. Em 2015, Moscetti et al., estudaram uma forma de classificar avelãs quanto a sua origem de acordo com a composição química. Os resultados foram obtidos usando dados de espectroscopia na região do infravermelho próximo e k-NN para discriminar as amostras de avelãs. O

resultado de exatidão durante a classificação foi satisfatório alcançando 95,5% de exatidão. Os autores concluíram que a metodologia adotada usando quimiometria foi uma forma rápida e fácil de se obter informações relevantes quanto a autenticação de espécies de avelãs.[41] No ano de 2016, Flood et al., avaliaram misturas de biodiesel no diesel a partir de dados de GC-MS. Foram concebidas duas análises exploratórias utilizando o k-NN e dados dos espectros, a primeira, objetivou discriminar as diferentes misturas amostrais, a depender da concentração, de biodiesel no diesel, a segunda análise, estava relacionada com a classificação de acordo com os tipos de biodiesel oriundos de matérias-primas diferentes, tais como: canola, soja, resíduos de gordura e sebo. Os resultados expressaram a utilidade do k-NN em conseguir classificar as duas análises realizadas. Como conclusão, abordaram a eficiência da técnica no tratamento de dados complexos, como os combustíveis, sem que houvesse a necessidade de pré-processamentos de dados.[42] Em 2019, Banerjee et al. discriminaram amostras de chá preto associando o k-NN aos dados sensoriais analisados. Os resultados mostraram que a eficiência desse método em classificar atingiu 99,75% de exatidão.[43]

Assim como o k-NN tem sido uma ferramenta útil em estudos de análise exploratória de dados químicos, no modo supervisionado, além do k-NN outros algoritmos lineares foram desenvolvidos, tais como: Análise de Discriminate Linear (LDA – *Linear Discriminant Analysis*), Análise Discriminante por Análise dos Componentes Principais (PCA-DA – *Principal component analysis discriminant analysis*) e Análise Discriminante por Mínimos Quadrados Parciais (PLS-DA – *Partial Least Squares-Discriminant Analysis*).

2.4 Análise de Discriminate Linear – LDA

A LDA, foi proposta por Ronald A. Fisher, em 1936, no intuito de resolver problemas associados a taxonomia vegetal. Fisher buscou um critério que maximizasse a separação entre duas ou mais classes de amostras levando em consideração as características peculiares. Para a execução deste critério é necessário definir uma superfície de decisão, denominado classificador. Em um espaço amostral, o classificador irá encontrar uma combinação linear, também conhecida como função discriminante, através da maximização da distância entre classes e minimização da distância intra-classe. Em outras palavras, o método

objetiva encontrar a melhor direção para um classificador de maneira que quando os dados são projetados em um plano, cada classe com as mesmas características possa ser separada.[9,44]

Na literatura são encontrados trabalhos de aplicação do LDA, alguns relacionados a matrizes de produtos naturais como o mel. A complexidade do mel dificulta muito a verificação da origem botânica, nesse contexto, a análise multivariada de dados tem desempenhado um papel fundamental na discriminação de amostras, facilitando a interpretação dos dados. Ciulu et al. (2018)[45] discriminam tipos de mel na Itália, quanto a origem relacionada as quatro categorias botânicas existentes. Para a análise discriminatória foram utilizados dados de sete propriedades físico-quimicas das 160 amostras de mel e a ferramenta de LDA. Os resultados de classificação obtidos apresentaram níveis altos de previsão, alcançando 100% de exatidão na validação externa. Concluíram que com o auxílio do LDA permitiu um controle rápido e fácil das amostras de mel quanto sua origem botânica. Essa mesma linha de pesquisa foi trabalhada por Oroian et al. (2018)[46] os resultados de classificação atingiram 94,12% de exatidão, mostrando que o método foi adequado para a autenticação botânica.

O modelo de LDA, apesar de ser muito empregado em análise de dados químicos, não é hoje a melhor opção para tratamento de dados instrumentais, particularmente aqueles com elevado número de variáveis, por exemplo, dados espectrais. Como desvantagem, a LDA pode ser aplicada apenas quando o número de variáveis é menor que o número de amostras, isso é decorrente das funções discriminantes lineares obtidas para classificar as amostras a partir de uma matriz de covariância que não possui a inversa.[47,48] Para evitar essa problemática, a PCA tem sido aplicada para a redução e seleção de dados anteriormente ao uso da LDA em análise exploratória de dados. Essa mistura de técnicas quimiométricas deu origem a um “novo” e poderoso método capaz de discriminar dados. A esse novo método em que se aplica a LDA no subespaço (scores) obtido na PCA, foi denominado como PCA-DA.[49-51]

O modelo de PCA-DA também tem sido uma importante ferramenta no estudo da matriz de mel, os autores anteriormente citados, Ciulu et al. (2018)[45], e Oroian et al. (2018)[46] publicaram artigos que se basearam em métodos validados de análises sensoriais e propriedades do mel que exigem um maior tempo para que o estudo seja

consolidado. Em busca da simplificação do tempo de análise do procedimento analítico, Zappi et al. (2018)[52] utilizaram *headspace* associada à cromatografia gasosa, em contrapartida, o número de variáveis obtidos em estudo foi superior, nesse caso, foi necessário aplicar a PCA, para se obter um conjunto reduzido de variáveis, seguido da LDA para discriminar os dados. Os resultados foram bons durante a validação, o conjunto teste apresentou valores superiores a 80% de exatidão para todas as classes, mostrando uma boa concordância entre essa nova abordagem e a tradicional para a determinação da origem botânica do mel.

Recentemente, estudos na área da saúde têm sido desenvolvidos associando Espectroscopia Raman (*Raman Spectroscopy*), para a identificação de biomoléculas relacionadas às doenças presentes no sangue, com a PCA-DA. As biomoléculas funcionam como marcadores químicos e a partir de uma análise espectral é possível obter informação de sua origem. Naseer et al. (2019)[53] investigaram o potencial das técnicas de Raman e PCA-DA para diferenciar tipos de condições patológicas causadas pela dengue e pela febre tifoide, uma vez que as sintomatologias das duas doenças são semelhantes. Essa nova metodologia foi proposta em resposta aos métodos tradicionais que possuem suas limitações, que incluem atraso de diagnóstico, falsos-positivos e suscetibilidade de infecção, além de ser necessário técnico bem treinado. Os autores alcançaram resultados impressionantes na classificação dos dois tipos de doenças. Concluíram, que a fusão PCA e LDA melhora bastante o desempenho de discriminação, em virtude de uma técnica extrair informações relevantes e favorecer a classificação pela outra técnica com maior eficiência.

De forma similar ao PCA-DA, foi desenvolvida outra técnica de classificação denominada PLS-DA através da junção de duas técnicas.[54] A PLS-DA surgiu há aproximadamente 20 anos da associação do método de Mínimos Quadrados Parciais (PLS - *Partial Least Squares*) e LDA.[55] Durante a construção do modelo PLS-DA, o objetivo principal é encontrar o número de variáveis latentes que descrevam a covariância nas amostras das matrizes e que tenham a correlação máxima com a classe de valores conhecidos, gerando menor peso para a classe irrelevante ou a variância do ruído.[56] A quantidade de variáveis latentes é estimada a partir da validação das amostras do conjunto de treinamento pelo método de validação cruzada como ocorre para os modelos PLS. Portanto, para cada variável latente a eficiência

da validação é expressa pelo Erro Quadrático Médio de Validação Cruzada (RMSECV - *Root Mean Squared Error Cross Validation*).[57,58] Como vantagem do método, após a construção do modelo PLS-DA, é possível obter a análise da Variável Importante na Projeção (VIP – *Variable Importance in Projec*tion) com intuito de identificar quais variáveis da técnica analítica utilizada foram mais relevantes para separação das classes das amostras.[59,60]

O método PLS-DA tem sido frequentemente incorporado na maioria das rotinas usadas pelos quimiometristas e tem sido considerada a ferramenta mais praticada em classificação de dados multivariados. Essa abordagem tem sido extensivamente aplicada em áreas relacionadas com a metabolômica,[60] proteômica,[61] genômica[62,63] e em áreas que geram grandes quantidades de dados como espectroscopia.[64] O crescente interesse no PLS-DA, especialmente no campo da metabolômica, deve-se principalmente à sua disponibilidade nos principais pacotes de software estatísticos como: R, S-Plus, SAS, SPSS e MATLAB.[64,65]

Estudos metabolômicos executados por Baddini et al. (2014),[66] com o objetivo de discriminar duas espécies de algas pardas *Dictyota ciliolata* e *Dictyota crenulata* a partir de dados de RMN de ^{1}H (Ressonância Magnética Nuclear de Hidrogênio) utilizaram a ferramenta de PLS-DA para discriminar as espécies foi empregada. Como resultado a ferramenta eficientemente classificou as amostras das duas espécies assim como possibilitou a aquisição das variáveis que mais influenciaram a separação, os esteróis e diterpenos. Outras aplicações estão relacionadas ao controle de qualidade de produtos, Stupak et al. (2018)[67] analisou a autenticidade de uísques usando amostras puras (malte) e misturas, os dados foram adquiridos por Microextração em fase Sólida acoplada a Cromatografia Gasosa com Espectrometria de Massas (SPME-GC-MS - *Solid-phase Microextraction Followed by Gas Chromatography-Mass Spectrometry*). Com o uso da PLS-DA foi possível discriminar as amostras e encontrar marcadores específicos que distinguiram uísques puros de falsos. Chen et al. (2019)[68] também utilizaram o PLS-DA como ferramenta de controle de qualidade. O estudo se baseou em fibras de quatro tipos de animais utilizados em indústrias têxteis. As amostras foram analisadas por infravermelho próximo, os dados gerados foram a base de informação para a classificação das fibras. Os resultados mostraram 100% sensibilidade e especificidade para as classes, indicando que a

associação da técnica de infravermelho e PLS-DA são viáveis para esse tipo de classificação.

As aplicações da PLS-DA e demais métodos citados até o presente momento, com exceção do k-NN, estão voltadas para solucionar problemas de natureza linear (Figura 2.6a). Em dados de natureza não linear, o k-NN não alcançou notoriedade, em parte, devido aos métodos mais recentes e robustos que estão sendo desenvolvidos para tratar dados de tal natureza (Figura 2.6b). Nesse caso, independente da quantidade de dados alguns problemas são muito difíceis de serem resolvidos por um único classificador, a fronteira de decisão que separa os dados de diferentes classes pode ser muito complexa. Como estratégia para minimizar os erros, a classificação dos dados pode ser realizada meio de uma combinação apropriada dos múltiplos classificadores, também conhecidos como *ensembles* (do inglês, aprendizagem por agrupamento).[69,70] Apesar da fundamentação teórica sobre esse método ter ocorrido na década de 90, o primeiro relato de trabalho sobre o particionamento do espaço de características usando dois ou mais classificadores foi datado em 1979 por Dasarathy e Sheela. Em 1990, Hansen e Salamon mostraram que a generalização de uma técnica de classificação poderia melhorar usando *ensembles*. Entretanto, o conhecimento teórico e expansão dessa técnica ocorreu a partir de 1996 com trabalho conceitual publicado por Leo Breiman.[69,70] A idéia é que o espaço de dados seja particionado em sub-espaços e em cada porção seja aplicado um classificador. O resultado final é obtido pela combinação das respostas de todos os classificadores. Esse artifício produz modelos mais robustos e aumenta o desempenho de classificação caso fosse aplicado um único classificador. [69,70]

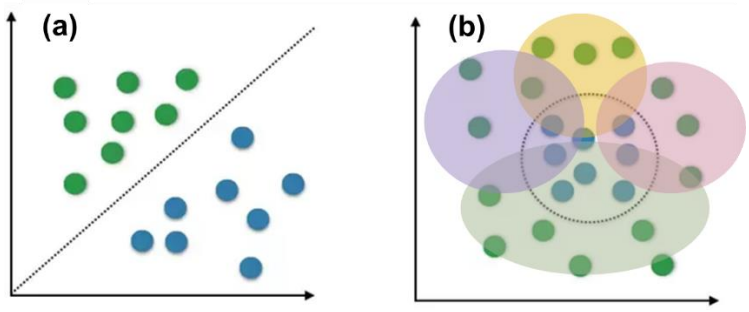

Figura 2.6 Sistema de classificação baseado em (a) único classificador e (b) múltiplos classificadores.

Nesse contexto, a partir de 1996 vários algoritmos *ensembles* passam a ser alvos de estudos, dentre eles, na vanguarda da quimiometria, Floresta Randômica

(RF), método de aprendizado por agrupamento capaz de tratar de dados complexos e não lineares.

2.5 Floresta Randômica – RF

RF, foi criada em 2001 pela fusão dos algoritmos Árvore de Classificação e Regressão (CART - *Classification and regression trees*) e do *ensemble* Agregação por *bootstrap* (bagging - *bootstrapping Aggregation*).[71] A primeira técnica citada, CART, está inserida na família de algoritmos baseados em árvores de decisão e suas primeiras abordagens ocorreram em 1966, nos Estados Unidos, com Hunt e outros pesquisadores formados em psicologia na tentativa de usar árvores de decisão para modelar o aprendizado de conceitos humanos.[72] Porém, foi em 1977 que Breiman, Friedman e colegas estatísticos aprimoraram a técnica e aplicaram a árvore em problemas de regressão e classificação. Em 1984, publicaram o artigo contendo a base do algoritmo CART, tecnicamente conhecido como partição recursiva binária.

No CART, em problemas qualitativos, uma única árvore de classificação é construída com nós conectados hierarquicamente, cada nó corresponde a um valor que está relacionado com a variável que contém a informação da classe. A partir dessa variável é realizado um corte (*split*) no espaço original de dados gerando sub-espaços, esse processo marca o início do crescimento da árvore. Assim, o crescimento da árvore (Figura 2.7) ocorre a partir do nó raiz, que possui a variável de maior contribuição para a discriminação, seguindo por partições recursivas resultando em ramos e nós descendentes (ou nós internos) até que se alcance uma resposta que permita a classificação, sendo este o nó decisão (folha) da árvore.[73-76] Para exemplificar o CART, o Anexo A, apresenta uma aplicação do algoritmo.

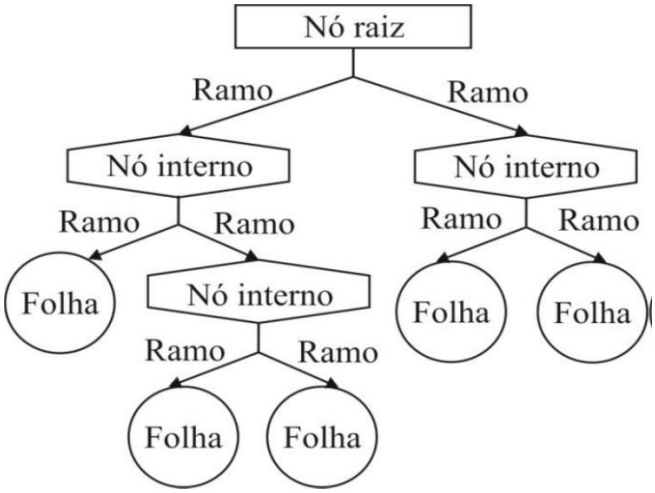

Figura 2.7 Esquema de uma árvore de decisão.

Para a decisão da variável preditora em cada nó são atribuídos critérios de partição binária, o mais conhecido é o índice de impureza de Gini ($Gini_{nó.}$). Esse índice é definido pela Equação (2.4):

$$Gini(nó) = 1 - \sum_{i=1}^{c} \left[p\left(\frac{i}{nó}\right)\right]^2 \qquad (2.4)$$

em que $p\left(\frac{i}{nó}\right)$ é a fração dos registros pertencentes à classe i no nó, e c é o número de classes.[75,76]

A diferença do índice de Gini entre o nó ascendente e nó descendente fornece um valor de ganho de informação. Isso resulta na escolha da variável preditora, que apresentou maior ganho de informação (Equação 2.5).

$$Ganho_{Gini} = Gini\ (ascendente) - \sum_{nd=1}^{n} \left[\left(\frac{nd}{na}\right) * Gini(nd)\right] \qquad (2.5)$$

onde n é o número de nós descendentes, nd é o número total de eventos do nó descendente e na é o número total de eventos do nó ascendente.[75,76]

Esse processo de divisões é interrompido quando a contagem de amostras no espaço é menor que a limitação mínima da quantidade de folhas predefinida; quando todas as amostras no subspaço representam a mesma classe.[77] Nesse contexto, as partições recursivas pelo índice de Gini podem se estender durante a construção da árvore até o ponto de classificar individualmente as amostras do conjunto de treinamento. A este ponto, podemos reduzir a capacidade de generalização do modelo comprometendo a exatidão quando o mesmo for aplicado a um novo conjunto de teste, esse fenômeno é denominado *overfitting*. Para contornar esse problema tem-se aplicado na RF a combinação de várias árvores como classificadores independentes

diminuindo o risco de decisões tendenciosas e *overfitting*. Neste caso, o processo de crescimento da árvore ocorre até atingir um certo grau de impureza, logo, é interrompido por um critério de parada de partição.[77,78] Outras desvantagens em se utilizar uma única árvore de decisão como classificador estão associadas a: instabilidade, isto é, uma pequena mudança no conjunto treinamento pode gerar uma significativa mudança na estrutura da árvore; dados espectroscópicos que apresentam alto grau de colinearidade podem comprometer a habilidade de predição do classificador, assim, *ensembles*, como RF, fazem uso de multiclassificadores solucionando esse problema. [77,79]

A classificação supervisionada com RF relaciona um conjunto de dados com um vetor classe (**y**) contendo a informação química qualitativa, amostrado de forma independente e com mesma distribuição para todas as árvores na floresta.[71] Neste processo (Figura 2.8), a partir do conjunto original (amostras e variáveis) a construção da primeira árvore inicia com a seleção randômica de amostras e variáveis. Na literatura é comum o uso da raíz quadrada do número de variáveis como critério inicial para escolher o número de variáveis para construção da árvore.[80-82] Outro fator importante está relacionado a segunda técnica que originou RF, *Bootstrap aggregation* (*Bagging*) (Figura 2.9).

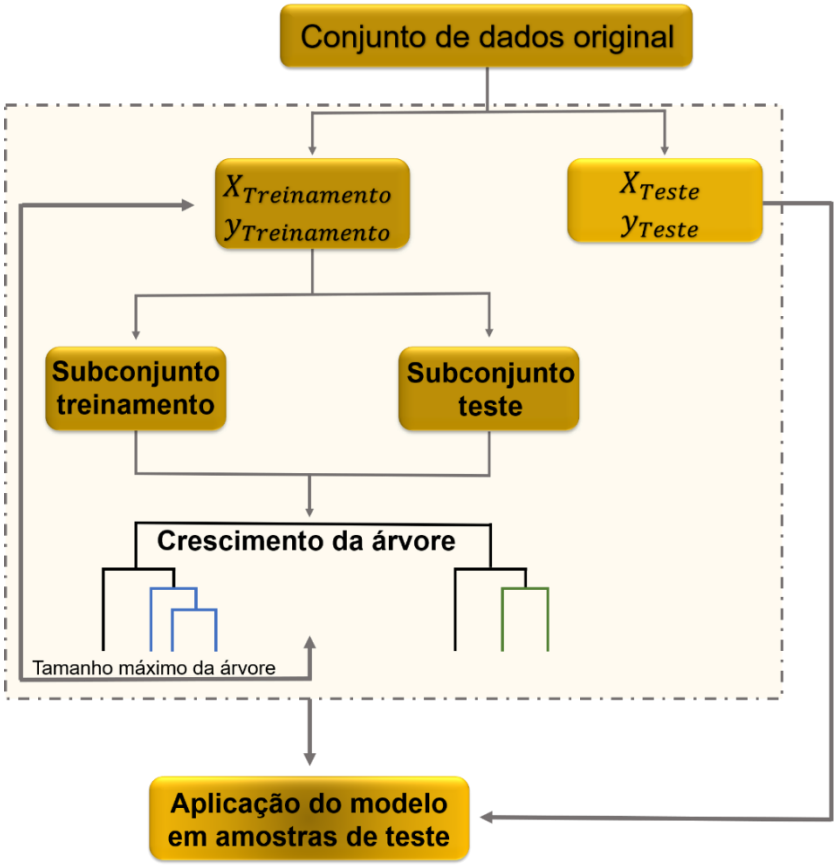

Figura 2.8 Fluxograma representativo das etapas seguidas em Floresta Randômica.

Esse *ensemble*, proposto por Breiman, em 1996, alavancou os demais métodos de aprendizagem por agrupamento.[69] O *bagging* tem como princípio a reamostragem com reposição para compor o conjunto treinamento das árvores de decisão, esse recurso foi também utilizado por Breiman em RF. Este modo de seleção de amostras permite criar diferentes sub-conjuntos amostrais para compor cada árvore de decisão. Essa estratégia amplia o ranking da matriz de amostras, o que é importante quando se trabalha com uma grande quantidade de variáveis e poucas amostras.[83] Entretanto, Granitto et al. (2007)[84], Gromski et al. (2015)[65] e Kucheryavskly (2018)[79] destacam que a principal vantagem de RF em relação ao *bagging* está na seleção randômica de variáveis que fundamenta apenas a técnica mais recente.

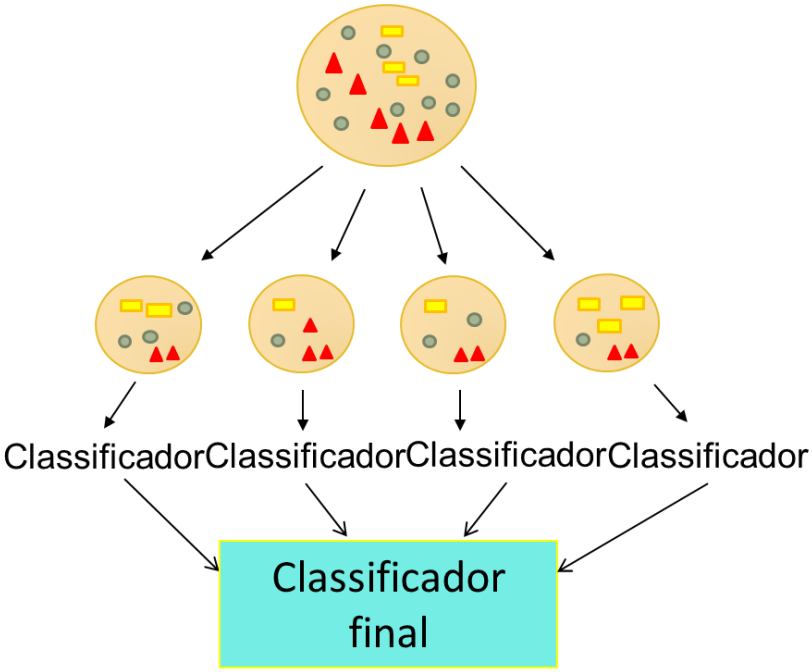

Figura 2.9 Esquema ilustrativo de reamostragem com reposição.

Em RF, partindo dos dados originais é o obtido o conjunto de treinamento, chamado *in bag*, (aproximadamente 2/3 do total de amostras) e conjunto teste, chamado *out-of-bag* (aproximadamente 1/3 do total de amostras).[71,85-87] Durante a etapa *in bag* (Figura 2.8), o conjunto treinamento é dividido em k subconjuntos de treinamento e teste, mutuamente exclusivos de mesmo tamanho, utilizando o método de *k-fold* durante a validação cruzada. Nesta etapa o subconjunto de teste é utilizado para a validação enquanto os k-1 restantes são utilizados para a construção do modelo. O procedimento é repetido k vezes até que todos os subgrupos tenham sido utilizados como validação.[9] Concomitante à separação dos subconjuntos, variáveis preditoras são identificadas dentro da seleção randômica pelo critério de Gini e são utilizadas para compor os nós durante a construção da árvore de classificação. Em seguida, é possível alimentar essa árvore com o conjunto teste (*out of bag*), que não participou de sua construção, a fim de classificá-las e verificar a exatidão (desempenho) do modelo criado para esse novo conjunto de dados.[74,87]

Todo esse processo de construção de uma única árvore é repetido iterativamente até que se conclua o número total de árvores (N) (Figura 2.8).[78] Na literatura o número total de árvores preditoras é variável, há relatos de 300,[89] 500, [81,90]

1000[91] e 2000 árvores.[92] Essa variabilidade do número de árvores pode estar associada à complexidade de informação na matriz original de dados, recentemente, Khoury et al. (2019)[93] desenvolveram estudos na Medicina para diferenciar pacientes diagnosticados com esclerose múltipla ou esclerose lateral amiotrófica. Os dados de soro dos pacientes doentes e saudáveis foram analisados por infravermelho e discriminados por 100 árvores preditoras de RF. Os resultados mostraram 100% de especificidade para os pacientes portadores da doença de esclerose múltipla e 98% para pacientes portadores da esclerose lateral amiotrófica. Xu et al. (2019)[94] utilizam dados de infravermelho e RF para classificar amostras da espécie *Angelica dahurica* de acordo com a origem. Nesse estudo, os autores realizaram estudos de otimização do número de árvores que mostrasse melhor desempenho na discriminação dos dados. Um total de 100 árvores foram utilizadas na construção de RF e os resultados foram satisfatórios. A mesma técnica espectroscópica foi utilizada por Santana et al. (2019)[95] para discriminar com RF amostras de óleos adulterados. Nesse trabalho, diferentemente de Khoury et al. (2019)[93] e Xu et al. (2019)[94] foi necessário utilizar um maior número de árvores de classificação (500) nos dados de infravermelho para que os resultados fossem promissores.

A literatura descreve controvérsias de autores sobre o número ótimo de árvores de classificação. Belgiu e Dragut (2016)[96] publicaram um artigo de revisão relatando que a maioria dos trabalhos citados no artigo definiram o número de árvores como 500, uma vez que os erros eram estabilizados antes mesmo que o número máximo de árvores de classificação fosse alcançado. Outro motivo foi apontado pelos autores para que essa quantidade de árvores fosse comumente usada e estaria associado a disponibilidade de RF no pacote R tendo esse valor como padrão já pré-estabelecido pelo algoritmo. Em 2018, um estudo mais aprofundado sobre a otimização do número de árvores foi proposto por Kucheryavskly (2018),[79] o artigo afirmou que utilizar 1000 – 2000 árvores é eficiente e suficiente independente do tamanho da quantidade de dados originais por produzir modelos com variação randômica relativamente pequena. A escolha dentro da faixa mencionada favorece a construção de árvores em menor tempo, dependendo da capacidade computacional disponível. Em comparação a faixa abordada anteriormente, exemplos de aplicação mostraram que a construção de RF com maior número de árvores (por exemplo, 10.000 a 20.000) tornou a variação

randômica insignificante, entretanto, não influenciou no desempenho do modelo e como desvantagem demandaria maior tempo computacional.

O uso dessas várias árvores de decisão, ou seja, a predição com a utilização de vários classificadores independentes é vantajosa por ter distribuída a variância do conjunto de dados.[69,97] Cada classificador individual (CART) contribui com um voto final sobre a classe predita para o conjunto teste, essas decisões independentes são combinadas em um único classificador que resulta na classe com maior número de votos.[97,98]

A literatura apresenta aplicações de RF em classificação supervisionada em diversas matrizes contendo dados lineares ou não lineares. Os resultados dessas aplicações têm confirmado a potencialidade de RF em problemas de classificação quando associada, principalmente, aos métodos analíticos que permitem aquisição de grande quantidade de informações como: espectroscópia de infravermelho, raman, RMN de ^{1}H, RMN de ^{13}C, espectrometria de massas e cromatografia.[86,99,100] O gráfico (Figura 2.10), apresenta a evolução do número de publicações sobre o algoritmo RF e suas aplicações nos últimos 20 anos.[101]

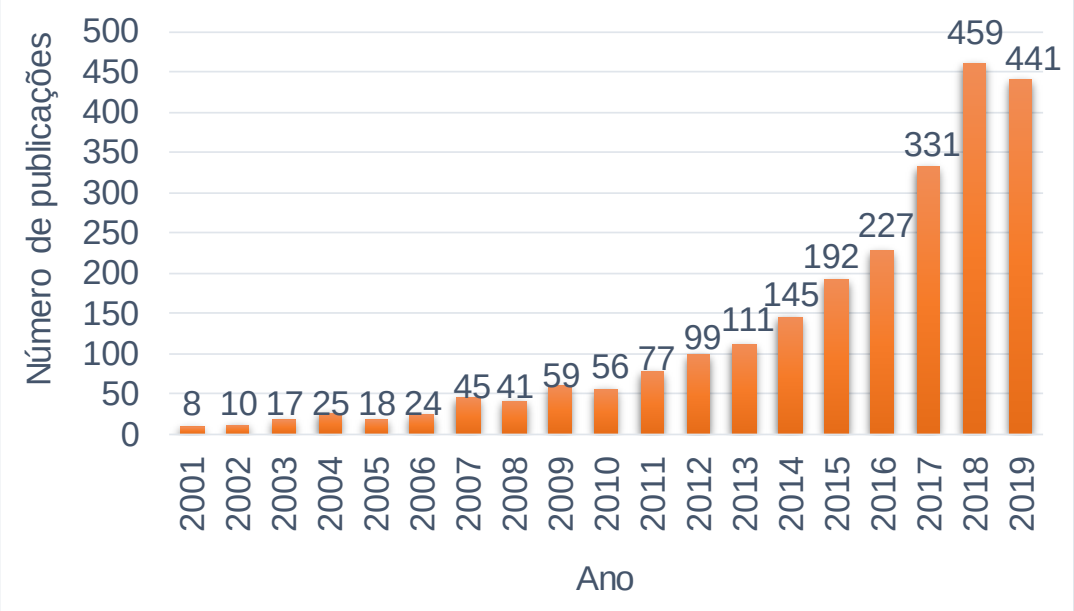

Figura 2.10 Número de publicações de RF em 20 anos nas áreas de química. Tópico (palavra-chave): Random Forest.[101]

A pesquisa mostra que ao longo dos anos houve um crescimento significativo na aplicação dessa técnica no tratamento de dados químicos, totalizando 2365 publicações desde quando RF foi criada. Nos últimos 4 anos, a média anual foi de 364 trabalhos aplicando RF, especialmente, nas áreas de análise genética, metabolômica, análise de imagens médicas, controle de qualidade e autenticação de alimentos.[86,99,100] Entretanto, a acelerada modernização na área de controle de qualidade e autenticação de alimentos e consequentemente, maior aporte de informação, possibilitou uma demanda maior do uso de RF. Nesse ramo, Carvelo et al. (2019)[70] reportam que a aplicação de RF, desde 2018, tem expandido em maiores proporções quando comparada a outros métodos atuais de aprendizagem de máquina como: CART e Máquina de Vetor Suporte (SVM - *Support Vector Machine*). Os autores justificam que esse aumento de trabalhos é decorrente da habilidade que RF possui em solucionar problemas que envolvem múltiplas classes.

A maioria dos trabalhos publicados abordam a eficiência da aplicação de RF a partir do algoritmo original, criado por Breiman,[71,72] em diversas matrizes, sem maior descrição sobre os cálculos internos realizados durante a etapa de aprendizagem (construção das árvores), dentre as aplicações: Liu et al. (2016)[91] avaliaram o poder discriminatório de diferentes técnicas quimiométricas, incluindo RF, em sementes de arroz transgênicas e não transgênicas. A maior capacidade preditiva foi atribuída a técnica de RF com 96,67% de exatidão. Essa mesma matriz é objeto de estudo de Maione et al. (2016)[102] os autores utilizaram diferentes métodos dentre eles RF para discriminar dados de ICP-MS de amostras de arroz quanto sua origem geográfica. Os resultados de RF para exatidão, sensibilidade e sensitividade foram melhores em relação aos demais métodos.

Zhang et al. (2016)[98] estudaram a capacidade de classificação de três métodos quimiométricos, incluindo RF em amostras de aço com composições distintas, a partir de dados obtidos por Espectroscopia de Emissão em Plasma induzido por Laser (LIBS - *Laser-induced Breakdown Spectrocopy*). Os resultados mostraram que a técnica de RF foi mais promissora que as demais na discriminação do conjunto amostral. Ismael et al. (2017)[103] utilizaram PLS-DA e RF para classificar o óleo resinoso obtido da *Aquilaria malaccensis Lamk* em um estudo metabolômico com dados obtidos por RMN de ^{1}H. Os resultados mostraram que com RF além de classificar de acordo com a qualidade do óleo nos diferentes níveis, confirmou a classificação obtida pelo método

tradicional utilizado pelos comerciantes desse produto. Canizo et al. (2018)[100] classificaram amostras de sementes de uvas de diferentes regiões de Mendonza (Argentina) com dados de metais obtidos por ICP-MS e diferentes técnicas quimiométricas com a finalidade de comparar seus desempenhos na discriminação dos grupos. Os autores concluíram que o método de RF mostrou um melhor desempenho comparado aos demais métodos com valor de exatidão de 97%.

Esses trabalhos mostram a capacidade de RF em classificar quando é fornecido um aporte maior de dados. Entretanto, a depender do método analítico, em especial os espectroscópicos, utilizados na obtenção de informações de amostras complexas, a quantidade de variáveis pode ser muito elevada, acarretando um elevado custo computacional, além de demandar maior tempo durante a modelagem.[104-106] Nesse sentido, para obter modelos mais rápidos e parcimoniosos torna-se importante o uso de métodos que selecionem as variáveis de maior relevância, anteriormente a aplicação de RF. Esses métodos são capazes de proporcionar uma redução da dimensionalidade de dados e podem destacar as variações que diferenciam as informações entre os espectros.[106]

A literatura é escassa de trabalhos que abordem estratégias de seleção ou redução de variáveis, dentre elas a PCA, em dados espectroscópicos combinadas a RF. Trabalhos que usam a PCA como método de seleção de variáveis, são obtidos os escores contendo a informação mais relevante. Em seguida, a matriz de escores é utilizada como parâmetro de entrada para os modelos de classificação com RF. Shen et al. (2018),[107] utilizam essa metodologia para discriminar alfaces contaminadas com diferentes concentrações de cádmio (Cd) a partir da LIBS. Os autores classificaram as informações por duas formas: a primeira usando somente RF, e a segunda, associando a PCA ao RF. O resultado obtido apenas com RF foi de 80% de exatidão, considerado baixo quando comparado com a fusão da PCA e RF. A segunda forma mostrou 100% de exatidão na classificação das amostras de alface. Essa mesma ideia, foi aplicada por Amjad et al. (2018),[108] para discriminar amostras de leite pela origem biológica: búfala, vaca, cabra e humano. Os espectros de raman dos tipos de leite demonstraram diferentes intensidades decorrentes das variações quanto ao teor de proteínas, lactose, gordura e outros nutrientes. Foram aplicados os métodos combinados de PCA e RF nos dados e os resultados alcançaram valores de 93,7% de exatidão para classificar o leite nas quatro classes.

Essas múltiplas aplicações de RF, sendo usada como a única técnica discriminante ou quando combinada a outro método para classificar matrizes com diferentes níveis de complexidade demonstram a multifuncionalidade e eficiência para atuar em reconhecimento de padrões. Entretanto, são escassos relatos na literatura de estudos voltados para o pré-processamento de dados, detalhamento do algoritmo, modificações no algoritmo original criado por Breiman,[71] identificação das variáveis com maior relevância utilizadas pelos classificadores de RF e associação de RF a outros métodos de seleção de variáveis. Selecionar variáveis que contribuam positivamente na classificação de dados poderiam potencializar o desempenho de RF. Além disso, ter conhecimento de quais variáveis estão influenciando na discriminação de um conjunto de dados seria um grande passo a ser dado na Química Analítica.

2.6 Parâmetros de avaliação qualitativos

Para avaliar o desempenho dos algoritmos classificadores se faz necessário o uso de métricas de avaliação para verificar se o resultado da predição é satisfatório. Esses parâmetros relacionam resultados de classes previstas pelo modelo com as classes referências. Partindo de um problema de classificação contendo duas classes, positivas e negativas, os resultados de previsão dos modelos de classificação podem ser avaliados pela sensibilidade, especificidade, exatidão. A sensibilidade (Equação 2.6) mostra a probabilidade da amostra ser prevista como classe positiva pelo modelo e ela realmente pertencer a classe positiva.

$$sensibilidade = \frac{TP}{TP+FN} \qquad (2.6)$$

em que TP representa o número de amostras positivas devidamente idendificadas pelo classificador e FN referem-se às amostras positivas que foram incorretamente identificadas pelo classificador. Logo, representam os falsos negativos. Modelos com alta sensibilidade implicam em FN ser o menor possível e ser correto na classificação.

Em contrapartida, a especificidade (Equação 2.7) é a probabilidade da amostra não ser prevista como classe positiva pelo modelo e ela realmente não pertencer a essa classe.

$$especificidade = \frac{TN}{TN+FP} \qquad (2.7)$$

em que TN referem-se às amostras negativas que foram corretamente identificadas pelo classificador e FP referem-se às amostras negativas que foram incorretamente Identificadas pelo classificador. Logo, representam os falsos positivos.

A exatidão (Equação 2.8) relaciona todos os resultados envolvidos na classificação e representa a porcentagem de amostra do conjunto de teste que foram classificadas corretamente, ou seja, é a taxa de acerto global.[109]

$$exatidão = \frac{TP+TN}{TP+TN+FP+FN} \qquad (2.8)$$

De forma geral, em Química Analítica, esses parâmetros avaliam a capacidade dos métodos quimiométricos em relacionar as informações contidas dos dados químicos obtidas por técnicas instrumentais, como a espectroscopia, com as propriedades de interesse da matriz em estudo. As técnicas espectroscópicas são capazes de fornecer informações a nível molecular. Entre várias técnicas espectroscópicas, a espectropia de RMN, em especial, apresenta uma boa resolução espectral, a qual apresenta ainda a vantagem de permitir a determinação não só qualitativa, mas também quantitativa, de uma larga faixa de tipos de amostras. Diferentes grupos funcionais aparecem em regiões espectrais bem resolvidos, e este fator tem estabelecido a RMN como uma técnica analítica altamente discriminatória e quantitativamente confiável.[110] Assim, métodos de análise utilizando a RMN têm despertado grande interesse a fim de otimizar a caracterização de amostras complexas como petróleos e derivados.[104,111,112]

2.7 Ressonância Magnética Nuclear (RMN) e petróleo

A espectroscopia RMN ocorre devido a interação da matéria com a radiação eletromagnética.[113] Os núcleos possuem diversas propriedades físicas destacando-se, massa, carga elétrica, momento magnético e *spin* nuclear. Sendo o momento magnético e spin nuclear fundamentais para a ocorrência do fenômeno da RMN, os quais conferem ao núcleo uma propriedade denominada magnetismo, isto é, confere ao núcleo um comportamento semelhante a pequenos imãs.[114]

Apesar de núcleos iguais apresentarem mesmo valor de razão magnetogírica, que é uma constante de cada isótopo, há ocorrência de diferentes valores de frequência, consequentemente diferentes deslocamentos químicos. Isto indica que

cada núcleo experimenta um campo magnético distinto, de acordo com sua posição na molécula. Diversas referências apresentam tabelas de deslocamentos químicos típicos encontrados para os principais núcleos estudados (^{1}H e ^{13}C). Em petróleo, os autores, Poveda e Molina (2012),[115] apresentaram estudo contendo os deslocamentos de acordo com os assinalamentos estruturais de hidrogênios (Anexo C, Tabela C.1) e carbonos (Anexo C, Tabela C.2) para essa matriz e seus derivados. Essas informações químicas obtidas pela técnica de RMN tem mostrado potencial na identificação e quantificação de diferentes tipos de hidrocarbonetos e pode fornecer informações sobre o conteúdo relativo de componentes alifáticos (0,1 – 4,5 ppm para hidrogênio e 3 – 78 ppm para carbono), olefínicos (4,5 – 6,0 ppm para hidrogênio e 85 – 129,2 ppm para carbono) e aromáticos (6-9,0 ppm para hidrogênio e 129,2 -137 para carbono).

Assim, a RMN tem sido aplicada em estudos sobre constituintes do petróleo a fim de relacionar essa informação a diferentes propriedades físico-químicas do óleo. Entretanto, os espectros de RMN do petróleo bruto e seus derivados são muito complexos e contêm um grande número de sinais refletindo sua grande complexidade química. Portanto, para explorar o conteúdo de todas as informações dos dados de RMN, em petróleos, vários métodos de análise multivariada têm sido desenvolvidos e trabalhados em conjunto com essa técnica.[103-106,111,116] Neste trabalho, a RMN foi associada a RF para classificação de petróleos em relação ao número de acidez total e ponto de fluidez máximo. Através dessas informações e recursos de RF também foi possível a identificação das variáveis que mais contribuíram para a classificação dos dados.

2.8 Referências Bibliográficas

1. FERREIRA, M.; ANTUNES, A. M.; MELGO, M. S.; VOLPE, P. L. O. Quimiometria I: calibração multivariada, um tutorial. *Química Nova,* **22,** 724-731, 1999.

2. KOWALSKI, B. R.; SEASHOLTZ, M. B. Recent developments in multivariate calibration. *Journal Chemometrics,* **5**, 129, 1991.

3. KIRALJ, R.; FERREIRA, M. M. C.The past, present, and future of chemometrics world wide: some etymological, linguistic and bibliometric investigations. *Journal Chemometrics*, **20**, 247–272, 2006.

4. KOWALSKI B. R. Chemometrics: views and propositions. *Journal of Chemical Information and Computer Science*, **5**, 201–203, 1975.

5. NETO, B, B.; SCARMINIO, I. S.; BRUNS, R. E. 25 anos de quimiometria no Brasil. *Química Nova*, **29**, 1401-1406, 2006.

6. WOLD, S.; SJÖSTRÖM, M. Chemometrics, present and future success. *Chemometrics and Intelligent Laboratory Systems,* **44**, 3-14, 1998.

7. CORREIA, P. R. M.; FERREIRA, M. M. C. Reconhecimento de padrões por métodos não supervisionados: explorando procedimentos quimiométricos para tratamento de dados analíticos. *Química Nova,* **30**, 2, 481-487, 2007.

8. LAVINE, B. K. Chemometrics. *Analytical Chemistry*, **72**, 91-97, 2000.

9. BRERETON, R.G. Chemometrics Data Analysis for the Laboratory and Chemical Plant. New York: John Wiley & Sons Ltda, 2003,183p.

10. FISHER, R. A.; MACKENZIE, W. A. Studies in crop variation: II. The manurial response of different potato varieties. *Journal of Agricultural Science*, **13**, 311-320, 1923.

11. JOLLIFFE, I. T. Principal Component Analysis. Second ed. Springer Series in Statistics. New York: Springer-Verlag New York, 2002, 1p.

12. RERIS, R.; BROOKS, P. Principal Component Analysis and Optimization: A Tutorial*. Informs Computing Society*, **11**, 212–225, 2015.

13. FERREIRA, M. M. C. Quimiometria: conceitos, métodos e aplicações. 1ed. São Paulo: Editora Unicamp, 2015, 1p.

14. HOTELLING, H. Analysis of a compfex of statistical variables into principal components, *Journal of Educatjonaf Psychology*, **24**, 498-520, 1933.

15. WOLD, S.; ESBENSEN. K.; GELADI, P. Principal Component Analysis. *Chemometrics and Intelligent Laboratory Systems*, **2**, 37-52, 1987.

16. BRO R.; SMILDE A. K. Principal component analysis. *Analytical Methods*, **6**, 2812-31, 2014.

17. MILDNER-SZKUDLARZ, S.; JELEN, H. H. The potential of different tchiques for volatile compounds analyses coupled with PCA for the detection of adulteration of olive oil with hazelnut oil. *Food Chemistry*, **110**, 751-761, 2008.

18. LYRA, W. S.; SILVA, E. C.; ARAÚJO, M. C. U.; FRAGOSO, W. D.; VERAS, G. Classificação periódica: um exemplo didático para ensinar análise de componentes principais. *Química Nova,* **33,** 7, 2010**.**

19. SABINO, J. F. P.; REYES, M. M.; BARRERA, C. D. F.; SILVA, A, J, R. Análise e discriminação de quimiotipos de *Lippia graveolens* H.B.K. da Guatemala por microextração em fase sólida, CG-EM e Análise multivariada. *Química Nova*, **35**, 1, 97-101, 2012.

20. KHANMOHAMMADI, M.; GARMARUDI, A. B.; GUARDIA, M. Characterization of petroleum-based products by infrared spectroscopy and chemometrics. TRAC Trends in Analytical Chemistry, **35**, 135-49, 2012.

21. SOUZA, M. O.; RAINHA, K. P.; CASTRO, E. V. R.; CARNEIRO, M. T. W. D.; FERREIRA, R. Q. Análise exploratória das concentrações dos metais Na, Ca, Mg, Sr e Fe em extrato aquoso de petróleo, determinados por ICP OES, após otimização empregando planejamento de experimentos. *Química Nova*, **7**, 980-986, 2015.

22. SPEIGHT, J. G.; Handbook of Petroleum Analysis, John Wiley and Sons Inc: New Jersey, **2001**, 519p.

23. FARAH, M. A. Petróleo e Derivados I: Caracterização do Petróleo e Seus Produtos. Rio de Janeiro: Curso de Formação em Engenharia de Processamento de Petróleo. UERJ/PETROBRAS, 2006.

24. VIEIRA, L. V. V.; RAINHA, K. P.; CASTRO, E. V. R. FILGEUIRAS, P. R.; CARNEIRO, M. T. W. D.; BRANDÃO, G. P. Exploratory data analysis using API gravity and V and Ni contents to determine the origins of crude oil samples from petroleum fields in the Espírito Santo Basin (Brazil). *Microchemical Journal.* **124,** 26-30, 2016.

25. CHIABERGE, S.; FIORANI. T.; SAVIONI, A.; BIONDA, A.; RAMELLO, S.; CESTI, M. P. P. Classification of crude oil samples through statistical analysis of APPI FTICR mass spectra. *Fuel Processing Technology*, **106**, 181–185, 2013.

26. PAIVA, E. M.; ROHWEDDER, J. J. R.; PASQUINI, C.; PIMENTEL, M. F.; PEREIRA, C. F. Quantification of biodiesel and adulteration with vegetable oils in

diesel/biodiesel blends using portable near-infrared spectrometer. *Fuel*, **160**, 57–63, 2015.

27. ALTINPINAR, S.; SORAK, D.; SIESLER, H. W. Near infrared spectroscopic analysis of hydrocarbon contaminants in soil with a hand-held spectrometer. *Journal Near Infrared Spectrometry*, **21**, 511–521, 2013.

28. CORREIA, R. M.; TOSATO, F.; DOMINGOS, E.; RODRIGUES, R. R. T.; AQUINO, L. F. M.; FILGUEIRAS, P. R.; LACERDA, V. L. J.; ROMAO, W. Portable near infrared spectroscopy applied to quality control of Brazilian Coffee. *Talanta*, **176**, 59–68, 2018.

29. KEMP, C.; TENENBAUM, J. B. The discovery of structural form. *National Academy Sciences*, 10687-10692, 2008.

30. ADAMSON, G. W.; BAWDEN D. Comparison of Hierarchical Cluster Analysis Techniques for Automatic Classification of Chemical Structures. *Journal of Chemical Information and Computer Science*, **21**, 4, 204-209, 1981.

31. HOBERT, H.; MEYER, K. Infrared spectroscopic analysis of complex mixtures by classification and regression - Characterization of urinary calculi. *Fresenius Journal Analytical Chemical*, **344**, 78 -185, 1992.

32. BARBOSA, I. S.; BRITO, G. B. SANTOS, G. L. SANTOS, L. N. TEXEIRA, L. S. G.; ARAUJO, R. G. O.; KORN, G. A. Multivariate data analysis of trace elements in bivalve molluscs: Characterization and food safety evaluation. *Food Chemistry*, **273** 64–70, 2019.

33. LIMA, A. M. S.; SANTOS, L. O.; DAVID, J. M.; FERREIRA, S. L. C. Mineral content in mustard leaves according to the cooking method. *Food Chemistry*, **273**,172–177, 2019.

34. CEBI, N.; DOGAN, C. E.; MESE, A. E.; OZDEMIR, D.; ARICI, M.; SAGDIC, O. A rapid ATR-FTIR spectroscopic method for classification of gelatin gummy candies in relation to the gelatin source*. Food Chemistry*, **277**, 373–381, 2019.

35. KOWALSKI, B. R.; BENDER, C. Pattern Recognition. A powerful Approach to Interpreting Chemical Data. *Journal of the American Chemical Society*. **94**, 5632-5639, 1972.

36. KOWALSKI, B. R.; BENDER, C. F. The K-Nearest Neighbor Classification Rule (Pattern Recognition) Applied to Nuclear Magnetic Resonance Spectral Interpretation. *Analytical Chemistry*, **44**, 8, 1405-1411,1972.

37. PICHLER, M. A.; PERONE, S. P. Computerized Pattern Recognition Applications to Chemical Analysis Development of Interactive Feature Selection Methods for the K-Nearest Neighbor Technique. *Analytical Chemistry*, **46**, 12, 1790-1798,1974.

38. TSAI, C. W.; MILIAM, S. J.; TIPPLE, C. A. Exploring the analysis and differentiation of plastic explosives by comprehensive multidimensional gas chromatography-mass spectrometry (GC X GC–MS) with a statistical approach. *Forensic Chemistry*. **6**,10–18, 2017.

39. SLAVOV, S.; ALUSTA, P.; BUZATU, D. A.; WILKES, J. G. Feature selection from mass spectra of bacteria for serotyping Salmonella. *Journal of Analytical and Applied Pyrolysis*, **124**, 393–402, 2017.

40. PÉREZ-RODRÍGUEZ, M.; GAIAD, J. E.; HIDALGO, M. J. AVANZA, M. V.; PELLERANO, R. G. Classification of cowpea beans using multielemental fingerprinting combined with supervised learning. *Food Control*, **95**, 232–241, 2019.

41. MOSCETTI, R.; SAEYS, W.; KERESZTES, J. C.; GOODARZI, M.; CECCHINI, M.; DANILO, M., MASSANTINI, R. Hazelnut Quality Sorting Using High Dynamic Range Short-Wave Infrared Hyperspectral Imaging. *Food and Bioprocess Technology*, **8**, 1593-1604, 2015.

42. FLOOD, M.E.; CONOOLY, M. P.; COMISKEU, M. C., HUPP, A. M. Evaluation of single and multi-feedstock biodiesel – diesel blends using GCMS and chemometric methods. *Fuel*, **186**, 58–67, 2016.

43. BANERJEE, M. B.; ROY, R. B., TUDU, B.; BANDYOPADHYAY, R.; BHATTACHARYYA, N. Black tea classification employing feature fusion of E-Nose and E-Tongue responses. *Journal of Food Engineering*, **244**, 55–63, 2019.

44. FISHER, R. A. The use of multiple measurements in taxonomic problems. *Annals of human genetics***, 7**, 179-188, 1936.

45. CIULU, M.; SERRA, R.; CAREDDA, M.; SALIS, S.; FLORIS, I.; PILO, M. I.; SPANO, N.; PANZANELLI, A.; SANNA, G. Chemometric treatment of simple physical and chemical data for the discrimination of unifloral honeys. *Talanta,* **190**, 382–390, 2018.

46. OROIAN, M.; ROPCIUC, S.; PADURET, S. Honey authentication using rheological and physicochemical properties. *Journal of Food Science Technology*, **55**, 12, 4711-4718, 2018.

47. SWETS, D. L.; WENG, J. J. Using discriminant eigenfeatures for image retrieval. IEEE Transactions on Pattern *Analysis and Machine Intelligence*, **18**, 8, 831–836, 1996.

48. PEREIRA, R. C. C.; SKROBOT, V. L.; CASTRO, E. V. R.; FORTES, I. C. P.; PASA, V. M. D. Determination of Gasoline Adulteration by Principal Components Analysis-Linear Discriminant Analysis Applied to FTIR Spectra. *Energy & Fuels*, **20**, 1097-1102, 2006.

49. KEMSLEY, E. K. Discriminant analysis of high-dimensional data: a comparison of principal components analysis and partial least squares data reduction methods. *Chemometrics and Intelligent Laboratory Systems*. **33**, 1, 47-61, 1996.

50. SANDERCOCK. P. M. L.; PASQUIER, E. D. Chemical fingerprinting of unevaporated automotive gasoline samples. *Forensic Science International*, **134**, 1–10, 2003.

51. SKROBOT, V. L.; CASTRO, E. V. R.; PEREIRA, R. C. C.; PASA, V. M. D.; FORTES, I. C. P. Use of Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) in Gas Chromatographic (GC). Data in the Investigation of Gasoline Adulteration. *Energy & Fuels*, **21**, 3394–3400, 2007.

52. ZAPPI, A.; MELUCCI, D.; SCARAMAGLI, S.; ZELANO, A.; MARCAZZAN, G. L. Botanical traceability of unifloral honeys by chemometrics based on head-space gas chromatography. *European Food Research and Technology*, **244**, 2149–2157, 2018.

53. NASEER, K.; AMIN, A.; SALEEM, M.; QAZI, J. Raman spectroscopy based differentiation of typhoid and dengue fever in infected human sera. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, **206**, 97–201, 2019.

54. WOLD, S.; SJOSTROM, M.; ERIKSSON, L. PLS-regression: a basic tool of chemometrics, *Chemometrics and Intelligent Laboratory Systems*, **58**, 109–130, 2001.

55. BARKER, M.; RAYENS, W. Partial least squares for discrimination. *Journal of Chemometrics*, **17**, 166-173, 2003.

56. BALLABIO, D.; CONSONNI, V. Classification tools in chemistry. Part1: linear models. PLS-DA, *Analytical Methods*, **5**, 3790-3798, 2013.

57. BRERETON, R. G.; LLOYD, G. R. Partial least squares discriminant analysis: taking the magic away. *Journal of chemometrics*. **28**, 213-225, 2014.

58. ALMEIDA M, R.; CORREA, D. N; ROCHA, W. F. C.; SCAFI, F. J. O; POPPI, R. J. Discrimination between authentic and counterfeit banknotes using Raman

spectroscopy and PLS-DA with uncertainty estimation. *Microchemical Journal*, **109**, 170-177, 2013.

59. ALMEIDA, M. R.; FIDELIS, C. H. V.; BARATA, L. E. S.; POPPI, R. J. Classification of Amazonian rosewood essential oil by Raman spectroscopy and PLS-DA with reliability estimation. *Talanta*, **117**, 305-311, 2013.

60. SZYMANSKA, E.; SACCENTI, A. K.; SMILDE, J. A. Westerhuis, Double-check: validation of diagnostic statistics for PLS-DA models in metabolomics studies, *Metabolomics*, **8**, S3–S16, 2012.

61. CHRISTIN, C.; HOEFSLOOT, H. C, J.; SMILDE, A. K.; HOEKMAN, B.; SUITS, F.; BISCHOFF, R.; HORVATOVICH, P. A critical assessment of feature selection methods for biomarker discovery in clinical proteomics. *Molecular & Cellular Proteomics*, **12**, 263–276, 2013.

62. BOULESTEIX, A. L.; STRIMMER, K. Partial least squares: a versatile tool for the analysis of high-dimensional genomic data. *Briefings in Bioinformatics*, **8**, 32–44, 2007.

63. OKSMAN-CALDENTEY, K. M.; INZE, D. Plant cell factories in the post-genomic era: new ways to produce designer secondary metabolites. *Trends Plant Science*, **9**, 433–440, 2004.

64. BLEKHERMAN, G.; LAUBENBACHER, R.; CORTES, D. F.; MENDES, P.; TORTI, F. M.; AKMAN, S.; TORTI, S. V.; SHULAEV, V. Bioinformatics tools for cancer metabolomics. *Metabolomics*, **7**, 329–343, 2011.

65. GROMSKI, P. S.; MUHAMADALI, H.; ELLIS, D. I. XU, Y.; CORREA, E.; TUMER, M. L.; GOODACRE, R. A tutorial review: Metabolomics and partial least squares-discriminant analysis – a marriage of convenience or a shotgun wedding. *Analytica Chimica Acta.*, **879,**10–23, 2015.

66. BADDINI, A. L. Q.; NETTO, A.D. P.; TEIXEIRA, V. L. Metabolômica Baseada em RMN de ^{1}H e no Método dos Mínimos Quadrados Parciais Discriminante. *Revista Virtual de Química*, **6**, 2, 396-415, 2014.

67. STUPAK, M.; GOODALL, I.; TOMANIOVA, M.; PULKRABOVA, J.; HAJSLOVA, J. A novel approach to assess the quality and authenticity of ScotchWhisky based on gas chromatography coupled to high resolution mass spectrometry. *Analytica Chimica Acta*, **1042**, 60-70, 2018.

68. CHEN, H.; LIN, Z.; TAN, C. Classification of different animal fibers by near infrared spectroscopy and chemometric models. *Microchemical Journal*, **144**, 489–494, 2019.

69. BREIMAN, L. Bagging predictors. *Machine Learning*, **24**, 123-140, 1996.

70. CARVELO, A. M. J.; CASADO, A. G.; GONZALES, G. B. Alternative data minig/ machine learning methods for the analytical evaluation of food quality and authenticity – A review. *Food Research International*, **122**, 25-39, 2019.

71. BREIMAN, L. Random Forests. *Machine Learning*, **45**, 5-32, 2001.

72. QUINLAN, J. R. Induction of Decision Trees, Machine Learning. **1**,81-106, 1986.

73. ALSBERG, B. K.; GOODCRE, R.; ROWLAND, J. J.; KELL, D. B. Classification of pyrolysis mass spectra by fuzzy multivariate rule induction-comparison with regression, K-nearest neighbour, neural and decision-tree methods. *Analytica Chimica Acta*, **348**, 389-407, 1997.

74. CHO, J. H.; KURUP, P.U. Decision tree approach for classification and dimensionality reduction of electronic nose data. *Sensors and Actuators B: Chemical*, **160**, 542-548, 2011.

75. SHANG, W.; HUANG, H.; ZHU, H.; LIN, Y.; QU, Y.; WANG, Z. A novel feature selection algorithm for text categorization, *Expert Systems with Applications*. **33**, 2007, 1–5, 2007.

76. BASHIR, S.; QAMAR, U.; KHAN, F. H.; JAVED, M. Y. MV5: A Clinical Decision Support Framework for Heart Disease Prediction Using Majority Vote Based Classifier Ensemble, *Arabian Journal for Science and Engineering*, **39**, 7771–7783, 2014.

77. AO, Y.; LI, H.; ZHU, L.; ALI, S.; YANG, Z. Identifying channel sad-body from multiple seismic attibutes with an improved random forest algorithm. *J. of petroleum Sci. Eng.,* **173**, 781-792, 2019.

78. CAO, D. S.; HU, Q. N.; XU, Q. S.; YANG, Y. N.; ZHAO, J. C.; LU, H. M.; ZHANG, L. X.; LIANG, Y. Z. In silico classification of human maximum recommended daily dose based on modified random forest and substructure fingerprint. Analytica Chimica Acta, **692**, 50–56, 2011.

79. KUCHERYASVSKLY, S. Analysis of NIR using decisions trees and their ensembles. *Journal of Analysis and Testing*, **2**, 274-289, 2018.

80. SVETNIK, V.; LIAW, A.; TONG, C.; CULBERSON, J. C.; SHERIDAN, R. P.; FEUSTON, B. P. Random forest: a classification and regression tool for compound classification and QSAR modeling, *Journal of Chemical Information and Computer Sciences*, **43**, 1947–1958, 2003.

81. POLIANCHIK, D. E.; GRIGORO'EV, V. Y.; SANDAKOV, G. I.; YARKOV, A. V.; BACHURIN, S. O.; RAEVSKII, O. A. Binary Cassification of CNS and PNS drugs. *Pharmaceutical Chemistry Journal,* **50**, 801-804, 2016.

82. SANTANA, F. B.; SOUZA, A. M.; POPPI, R.J. Visible and near infrared spectroscopy coupled to random forest to quantify some soil quality parameters, *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, **191**, 454–462, 2018.

83. ZHENG, L.; WATSON, D. G.; JOHNSTON, B. F.; CLARK, R. L.; EDRADA-EBEL, R.; ELSEHERI, W. A chemometric study of chromatograms of tea extracts by correlation optimization warping in conjunction with PCA, support vector machines and random forest data modeling, *Analytica Chimica Acta*, **642**, 257–265, 2009.

84. GRANITTO, P. M.; GASPERI, F.; BIASIOLLI, F.; TRAINOTTI, E.; FURNANELLO, C. Modern data mining tools in descriptive sensory analysis: a case study with a random forest approach. *Food Quality and Preference*, **18**, 681-689, 2007.

85. BIAU, G. Analysis of a Random Forests Model. *Journal of Machine Learning Research*, **13**, 1063-1095, 2012

86. LEE, S.; CHOI, H.; CHA, K.; CHUNG, H. Random forest as a potential multivariate method for near-infrared (NIR) spectroscopic analysis of complex mixture samples: Gasoline and naphtha. *Microchemical Journal*, **110**, 739–748, 2013.

87. STROZIER, E. D.; MOONEY, D. D.; FRIEDENBERG, D. A.; KLUPINSKI, T. P.; TRIPLETT, C. A. Use of Comprehensive Two-Dimensional Gas Chromatography with Time-of-Flight Mass Spectrometric Detection and Random Forest Pattern Recognition Techniques for Classifying Chemical Threat Agents and Detecting Chemical Attribution Signatures, *Analytical Chemistry*, **88**, 7068−7075, 2016.

88. AFANADOR, N. L.; SMOLINSKA, A.; TRAN, T. N.; BLANCHET, L. Unsupervised random forest: a tutorial with case studies, *Journal of Chemometrics*. **30**, 232–241, 2016.

89. LIN, X.; SUN, L.; LI, Y.; GUO, Z.; LI, Y.; ZHONG, K.; WANG, Q.; LU, X.; YANG, Y.; XU, G. A random forest of combined features in the classification of cut tobacco based on gas chromatography fingerprinting. *Talanta*, 82, 1571–1575, 2010.

90. DONG, J.; GUO, W.; LIU, D. Discrimintion of "Hayward" Kiwifruits Treated with Forchlorfenuron at different concentrations using hyperspectral imaging technology. *Food Analytical Methods*,**19**, 477-486, 2017.

91. VIGNEAUA, E.; COURCOUX, P.; SYMONEAUX, R.; GUÉRIN, L.; VILLIÈRE, A. Random forests: A machine learning methodology to highlight the volatile organic compounds involved in olfactory perception, *Food Quality and Preference*, **68**,135–145, 2018.

92. LI, Y.; ZHANG, J. Y.; WANG, Y. Z. FT-MIR and NIR spectral data fusion: a synergetic strategy for the geographical traceability of Panax notoginseng. *Analytical and Bioanalytical Chemistry*, **410**, 91–103, 2018.

93. KHOURY, Y.; COLLONGUES, N.; SÉZE, J.; GULSARI, V.; MENSAH, C. P.; MARCOU, G.; VARNECK, A.; NYAGAN, A. G. M.; HELLWIG, P. Serum-based differentiation between multiple sclerosis and amyotrophic lateral sclerosis by Random Forest classification of FTIR spectra. *Analyst*, **144**, 4647–4652, 2019.

94. XU, Y.; LIU, L.; HUANG, M.; XU, N. High accuracy deternimation of *Angelica dahurica* origin based on near infrared spectroscopy and random forest pruning algorithm. *Journal of Near Infrared Spectroscopy*. **27**, 1-8, 2019.

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

96. BELGIU, M.; DRAGUT, L. Random forest in remote sensing: A review of applications and future directions, *ISPRS J. Photogramm Remote Sens*., **114**, 24–31, 2016.

97. GISLASON, P. O.; BENEDIKTSSON, J. A.; SVEINSSON, J. R. Random Forests for land cover classification, *Pattern Recognition Letters*, **27**,294–300, 2006.

98. ZHANG, T.; XIA, D.; TANG, H.; YANG, X.; LI, H. Classification of steel samples by laser-induced breakdown spectroscopy and random forest, *Chemometrics and Intelligent Laboratory Systems*, **157**, 196-201, 2016.

99. LIU, W.; LIU, C.; HU, X.; YANG, J.; ZHENG, L. Application of terahertz spectroscopy imaging for discrimination of transgenic rice seeds with chemometrics, *Food Chemistry*. **210**, 415–421, 2016.

100. CANIZO, B. V.; ESCUDERO, L. B.; PÉREZ, M. B.; PELLERANO, R. G. Intra-regional classification of grape seeds produced in Mendoza province (Argentina) by multi-elemental analysis and chemometrics tools. *Food Chemistry*, **242**, 272–278, 2018.

101. PESQUISA base de dados. Web of Science. Disponível em: http://wcs-webofknowledge.ez43.periodicos.capes.gov.br/. Acessado em: 24 de setembro de 2019.

102. MAIONE, C.; LEMOS, B. B.; CAMPIGLIA, A. D.; BARBOSA, J. F.; BARBOSA, R. M. Classification of geographic origino f rice by data mining and inductively coupled plasma mass spectrometry. *Computers and Eletronics in Agriculture*, **121**, 101-107, 2016.

103. ISMAIL, S. N.; MAULIDIANI, M.; AKHTAR, M. T.; ABAS, F.; ISMAIL, I. S.; KHATIB, A.; ALI, N. A. M.; SHAARI, K. Discriminative Analysis of Different Grades of Gaharu (Aquilaria malaccensis Lamk.) by 1H-NMR-Based Metabolomics Using PLS-DA and Random Forests Classification Models, *Molecules*, **22**, 1612-1625, 2017.

104. MASILI, A., PULIGHEDDU, S. SASSU, L. SCANO, P.; LAI, A. Prediction of physical-chemical properties of crude oils by ^{1}H NMR analysis of neat samples and chemometrics. *J. Chemometr*., **50,** 729-738, 2012.

105. FILGUEIRAS P. R.; TERRA, L.; CASTRO, E. V. R.; OLIVEIRA, L. M. S. L.; DIAS, J. C. M.; POPPI, R. J. Prediction of the distillation temperatures of crude oils using ^{1}H NMR and support vector regression with estimated confidence intervals. *Talanta*. **142**, 197-205, 2015.

106. FILGUEIRAS, P. R.; PORTELA, N. A.; SILVA, S. R. C.; CASTRO, E. V. R.; OLIVEIRA, L. M. S. L.; DIAS, J. C. M.; NETO, A. C.; ROMÃO, W.; POPPI, R. J. Determination of saturates, aromatics, and polars in crude oil by ^{13}C NMR and support vector regression with variable selection by genetic algorithm. *Energy Fuels*. **30,**1972-1978, 2016.

107. SHEN, T.; KONG, W.; LIU, F.; CHEN, Z.; YAO, J.; WANG, W.; PENG, J.; CHEN, H.; HE, Y. Rapid Determination of Cadmium Contamination in Lettuce Using Laser-Induced Breakdown Spectroscopy. *Molecules*, **23**, 2930-2945, 2018.

108. AMJAD, A.; ULLAH, R.; KHAN, S.; BILAL, M.; KHAN, A. Raman spectroscopy based analysis of milk using random forest classification. *Vibrational Spectroscopy*, **99**, 124–129, 2018.

109. CUNHA-NETO, A.; OLIVEIRA, E. C. S.; LACERDA-JUNIOR, V.; CASTRO, E. V. R.; ROMÃO, W.; SILVA, S. R. C.; PEREIRA, R. G.; STEN, T.; FILGUEIRAS, P. R.; POPPI, R. J. Quality control of ethanol fuel: assessment of adulteration with methanol using 1h nmr. *FUEL*, **135**, 387-392, 2014.

110. NIELSEN, K. E., DITTMER, J., MALMENDAL, A., NIELSEN, N. C.. Quantitative analysis of constituents in heavy fuel oil by 1h nuclear magnetic resonance (nmr) spectroscopy and multivariate data analysis. ENERGY & FUELS, **22**, 4070-4076, 2008.

111. PORTELA, N. A.; OLIVEIRA, E. C. S; NETO, A. C.; RODRIGUES, R. R. T.; SILVA, S. R. C.; CASTRO, E. V. R.; FILGUEIRAS, P. R.. Quantification of biodiesel in petroleum diesel by 1H NMR: Evaluation of univariate and multivariate approaches. *Fuel*. **166**, 12-18, 2016.

112. MOLINA V., D.; URIBE, U. N.; MURGICH, J..Correlations between SARA fractions and physicochemical properties with 1H NMR spectra of vacuum residues from Colombian crude oils. *Fuel*, **89**, 185-192, 2010.

113. GIL, V. M. S.; GERALDES, C. F. G. C.. Ressonância Magnética Nuclear: Fundamentos, Métodos e Aplicações. 2. Ed. Lisboa: Fundação Calouste Gulbenkian, 2002. 1004 p.

114. LEVITT, M. H. Spin Dynamics: Basics of Nuclear Magnetic Resonance. 2. Ed. Southampton: John Wiley & Sons Ltd, 2008. 713 p.

115. POVEDA, J. C.; MOLINA, D. R. Average molecular parameters of heavy crude oils and their fractions using NMR spectroscopy. Journal of Petroleum Science and Engineering, **84**, 1-7, 2012.

116. DUARTE, L. M.; FILGUEIRAS, P. R.; DIAS, J. C. M.; OLIVEIRA, L. M. S. L.; CASTRO, E. V. R.; OLIVEIRA, M. A. L. Study of Distillation Temperature Curves from Brazilian Crude Oil by 1 H Nuclear Magnetic Resonance Spectroscopy in Association with Partial Least Squares Regression. *Energy Fuels*, **31**, 3892-3897, 2017.

3. APLICAÇÃO 1: USO DE FLORESTA RANDÔMICA NA IDENTIFICAÇÃO DE VARIÁVEIS IMPORTANTES

3.1 Introdução

Os métodos de classificação supervisionada estão sendo amplamente aplicados em tratamento de dados químicos devido a capacidade de classificar e generalizar os modelos.[1-3] Entretanto, há limitações das técnicas quimiométricas tradicionais (análise discriminante por análise de componentes principais – PCA-DA, análise discriminante por mínimos quadrados parciais – PLS-DA e k vizinhos mais próximos – k-NN) em classificar dados contendo grande quantidade de variáveis e um reduzido conjunto de amostras, como também de natureza não linear.[4] Em contrapartida, métodos recentes de aprendizagem de máquina como Floresta Randômica (RF) têm mostrado eficiência superando essas problemáticas. [3,5,6]

RF foi criada em 2001 pela fusão dos algoritmos CART (Classification and Regression Trees) e *Bootstrapping aggregation* (Bagging).[5] Este método é baseado na combinação de múltiplas árvores preditoras (classificadores fracos), sendo cada uma delas construída a partir de uma seleção randômica de amostras e de variáveis.[7,8]

No CART, uma árvore de classificação é construída com nós conectados hierarquicamente, cada nó corresponde a um valor (corte) que está relacionado com a variável que contém a informação da classe. O crescimento da árvore ocorre a partir do nó raiz, que possui a variável de maior contribuição para a discriminação, seguindo por partições recursivas resultando em ramos e nós descendentes até que se alcance uma resposta que permita a classificação, sendo este o nó decisão (folha) da árvore.[9,10]

Para a decisão da variável preditora em cada nó é atribuído um critério de partição binária, que é o índice de impureza de Gini ($Gini_{nó}$.). Esse índice é definido pela Equação (3.1):

$$Gini(nó) = 1 - \sum_{i=1}^{c} \left[p\left(\frac{i}{nó}\right) \right]^2 \qquad (3.1)$$

em que $p\left(\frac{i}{nó}\right)$ é a fração dos registros pertencentes à classe i no nó, e c é o número de classes.[9,10]

A diferença do índice de Gini entre o nó ascendente e nó descendente fornece um valor de ganho de informação. Isso resulta na escolha da variável preditora, que apresentou maior ganho de informação (Equação 3.2).

$$Ganho_{Gini} = Gini\ (ascendente) - \sum_{nd=1}^{n}\left[\left(\frac{nd}{na}\right) * Gini(nd)\right] \quad (3.2)$$

onde n é o número de nós descendentes, nd é o número total de eventos do nó descendente e na é o número total de eventos do nó ascendente.[9,10]

As partições recursivas pelo índice de Gini podem se estender durante a construção da árvore até o ponto de classificar individualmente as amostras do conjunto de treinamento. A este ponto, pode haver a redução da capacidade de generalização do modelo comprometendo a exatidão quando o mesmo for aplicado a um novo conjunto de teste, esse fenômeno é denominado *overfitting*. Para contornar esse problema tem-se aplicado na RF a combinação de várias árvores como classificadores independentes diminuindo o risco de decisões tendenciosas e *overfitting*. Nesse caso, o processo de crescimento da árvore ocorre até atingir um certo grau de pureza, logo, é interrompido por um critério de parada de partição.[12]

Todo esse processo de construção de uma única árvore é repetido iterativamente até que se conclua uma quantidade de árvores que configura o modelo de RF. Dessa forma, a classificação supervisionada pela RF relaciona um conjunto de dados com um vetor classe ($\boldsymbol{y}$), amostrado de forma independente e com mesma distribuição para todas as árvores na floresta.[5]

A literatura apresenta aplicações de RF em classificação supervisionada em diversas matrizes: Liu et al. (2016),[6] avaliaram o poder discriminatório de diferentes técnicas quimiométricas, incluindo RF, em sementes de arroz transgênicas e não transgênicas. A maior capacidade preditiva foi atribuída a técnica de RF com 96,67% de exatidão. Zhang et al. (2016),[13] estudaram a capacidade de classificação de três métodos quimiométricos, incluindo RF em amostras de aço com composições distintas, a partir de dados obtidos por LIBS. Os resultados mostraram que a técnica de RF foi mais promissora que as demais na discriminação do conjunto amostral. Ismael et al. (2017),[14] utilizaram PLS-DA e RF para classificar o óleo resinoso obtido da *Aquilaria malaccensis Lamk* em um estudo metabolômico com dados obtidos por ressonância magnética nuclear de hidrogênio (RMN de ^{1}H). Os resultados mostraram que com RF além de classificar de acordo com a qualidade do óleo nos diferentes níveis, confirmando uma classificação dada pelo método tradicional utilizado pelos comerciantes desse produto. Canizo et al. (2018),[15] classificaram amostras de sementes de uvas de diferentes regiões de Mendonza (Argentina) com dados de metais obtidos por espectrometria de massa por plasma acoplado indutivamente (ICP-

MS) e diferentes técnicas quimiométricas com a finalidade de comparar seus desempenhos na discriminação dos grupos. Os autores concluíram que o método de RF mostrou um melhor desempenho comparado aos demais métodos com valor de exatidão de 97%. Esses trabalhos de classificação por RF mostram a eficiência através dos satisfatórios resultados de exatidão desta técnica em discriminar amostras de diferentes níveis de complexidade, entretanto, não há estudos voltados para petróleos. Ademais os estudos não exploram a relação entre as variáveis mais importantes e exatidão do modelo.

Essas e outras abordagens na literatura têm confirmado a potencialidade de RF em problemas de classificação quando associada aos métodos espectroscópicos de infravermelho, raman, RMN de ^{1}H, espectrometria de massas e cromatografia.[6,7,15] Entretanto, não há relatos na literatura de estudos voltados para identificação de variáveis com maior relevância na discriminação de dados químicos quando aplicado a RF.

3.2 Objetivo

Desenvolver uma metodologia para obter informações das variáveis mais importantes na construção de modelos Floresta Randômica. A metodologia desenvolvida foi aplicada na discriminação de amostras de petróleo em relação ao ponto de fluidez máximo (PFM) utilizando dados de RMN de ^{1}H e ressonância magnética nuclear de carbono (RMN de ^{13}C).

3.3 Metodologia

Neste trabalho, foram utilizadas 105 amostras de óleo bruto, variando de leves a pesados (faixa de gravidade API de 11,4 a 52,4), cujos espectros de RMN de ^{1}H e de ^{13}C foram adquiridos e medida a propriedade físico-química do ponto de fluidez. O ponto de fluidez foi medido de acordo com o método padrão ASTM D5853,[16] e suas medidas foram usadas para avaliar as possíveis correlações com os compostos no óleo cru. Para esta avaliação, utilizou-se os teores de Saturados, Aromáticos e Polares (análise SAP). Este ensaio de SAP foi realizado por uma coluna cromatográfica preparativa, conforme descrito por Filgueiras et al. (2016).[17]

As amostras foram classificadas com base nos resultados dos ensaios de ponto de fluidez. Foi atribuído um valor referência para o ponto de fluidez de -9 ºC para se

obter duas classes. Na indústria do petróleo é de interesse obter informações de óleos com ponto de fluidez com valores negativos. Assim, a escolha desse valor referência foi associado a alta reprodutibilidade desse método que é de 9 ºC,[16] a fim de garantir de forma mais conservadora a classificação dos óleos em ponto de fluidez positivos e negativos. Para as 43 amostras com valores de ponto de fluidez acima de -9 ºC, foram atribuídas a “Classe 1”; e, para as 62 amostras com valores de ponto de fluidez abaixo de -9 ºC, foram atribuídas a “Classe 2”. A partir dessas atribuições de classes, desenvolvemos o modelo de classificação supervisionada por RF, cujos vários modelos foram construídos com diferentes pré-processamentos para a avaliação do desempenho. Durante a construção do modelo, utilizou-se a validação cruzada k-fold (CV – *Cross-validation*), para obter o crescimento da árvore, que foi aplicada para calcular a extidão dos modelos criados. O modelo que apresentou melhor desempenho (exatidão) foi selecionado para identificar as variáveis com maior contribuição na classificação. Um algoritmo em Matlab® foi desenvolvido para calcular a frequência na qual as variáveis (deslocamento químico dos espectros de RMN) foram utilizadas na construção das árvores selecionadas.

3.3.1. Aquisição de amostras e espectros de RMN

As 105 amostras de óleo bruto utilizadas neste trabalho foram provenientes da bacia sedimentar da costa brasileira, com gravidade API variando de leve (52,4) a óleo pesado (11,4) e ponto de fluidez entre -54 ºC a 39 ºC. Vale ressaltar que as amostras foram previamente desidratadas, com teor de água menor que 0,5% (v / v).[18]

3.3.1.1 Aquisição dos espectros do RMN de ^{13}C

Os espectros de RMN de ^{13}C foram obtidos a 27 ºC, utilizando um instrumento Varian VNMRS-400 com um campo magnético de 9,4 T, 10 mm de Banda Larga ^{15}N-^{31}P (^{1}H-^{19}F). As amostras foram preparadas dissolvendo-se uma massa de óleo (correspondente a 40% do volume final da solução em clorofórmio deuterado, $CDCl_3$) contendo 0,05 mol·L^{-1} de acetilacetonato de cromo III, 3 mL de volume total. As condições instrumentais estabelecidas foram: frequência, 100,51 MHz para o núcleo ^{13}C; janela espectral, 25510,2 Hz; tempo de aquisição, 1,285 s; tempo de espera, 7 s; pulso, 90° (14,2 µs); número de transientes, 1000; e modo de desacoplamento, nny.

O desacoplamento foi desligado durante o pulso, durante o tempo de espera, e foi conectado apenas durante a aquisição de dados.

3.3.1.2 Aquisição dos espectros do RMN de ^{1}H

Os espectros de RMN de ^{1}H foram obtidos a 27 ºC, utilizando um instrumento Varian VNMRS-400 com um campo magnético de 9,4 T, 5 mm de Banda Larga ^{1}H - ^{19}F. As amostras foram preparadas dissolvendo uma massa de óleo (correspondendo a 5% do volume final da solução em diclorometano, CH_2Cl_2), 600 µL do volume total. As condições instrumentais estabelecidas foram: frequência, 399,73 MHz; janela espectral, 6410,3 Hz; tempo de aquisição, 2,556 s; tempo de espera, 1 s; pulso, 90° (11,7 µs); número de transientes, 64.

3.3.2 Medidas do Ponto de Fluidez Máximo

A ASTM D5853[16] é específica para medições do ponto de fluidez máximo (superior) e mínimo (inferior) em óleos crus. O ensaio para determinar o ponto de fluidez foi realizado como descrito por Sad et al. (2014)[18] de acordo com a norma ASTM D5853. Neste método, as amostras foram homogeneizadas por 2 min e uma alíquota de 40 mg foi transferida para um frasco térmico acoplado ao termômetro. O teste do frasco, com a alíquota, foi mantido em repouso por 24 horas, na temperatura de 18 a 24 ºC. Após o período de repouso, o frasco de ensaio foi aquecido a 45 ºC por 15 min, e acoplado em uma câmera do equipamento semi-automático de teste de ponto de fluidez Pretrotest (R), modelo 531/5. Portanto, a temperatura foi diminuindo e um intervalo de 3 ºC foi o fluxo característico da amostra. Quando a amostra parou o fluxo, a temperatura foi registrada como o ponto de fluidez.

3.3.3 Quimiometria

3.3.3.1 Pré-processamento e processamento de dados

Antes do desenvolvimento do modelo de classificação, os espectros de RMN de ^{1}H e RMN de ^{13}C necessitavam ser alinhados para corrigir pequenos desvios nos deslocamentos químicos. Esse alinhamento reduz a variação de informações não relacionadas à composição química entre as amostras. Para o alinhamento foi aplicado o método de correlação por intervalos, com o algoritmo de icoshift.[19,20] As variáveis do espectro foram divididas em 50 intervalos iguais e foi realizado um deslocamento de cada intervalo, simultaneamente para todos os espectros do

conjunto de dados, com base em um espectro alvo, que pode ser um espectro médio. O algoritmo usa um mecanismo de otimização de correlação baseado na Transformada Rápida de Fourier (FFT - *Fast Fourier Transform*), para fornecer cálculos de velocidade e automação de alinhamento. Após o alinhamento dos espectros, foi realizado o pré-processamento dos dados, com avaliação dos métodos de pré-processamento mais utilizados: auto-escalamento, variação normal padrão (SNV), normalização, centralização de média e derivada de primeira ordem.[21-24]

3.3.3.2 Modelo de classificação de Floresta Randômica

A partir do conjunto de dados de RMN de ^{1}H (variáveis 20.433) e RMN de ^{13}C (65.000 variáveis) das 105 amostras de óleo bruto, desenvolvemos os modelos de classificação por RF (Figura 3.1). A construção da primeira árvore de classificação se inicia com a seleção randômica das amostras e variáveis. Na literatura é comum o uso da raiz quadrada no número de variáveis para indicar o número de variáveis selecionadas na construção de cada árvore,[25-28] porém, em virtude da grande quantidade de variáveis, neste trabalho, foram selecionadas randomicamente 10% do total de variáveis. Dessa maneira, uma parcela maior de variáveis pode ser selecionada durante o processo de construção das árvores. Isso possibilita que as variáveis mais significativas estejam presentes nessa seleção e, consequentemente, participem da construção de árvores. Outro fator importante é a seleção de amostras por *bootstrap* (reamostragem com reposição) para formar o conjunto treinamento, chamado in bag, (aproximadamente 2/3 do total, 70 amostras) e conjunto teste, chamado out-of-bag (aproximadamente 1/3 do total, 35 amostras). [5,7,8,28,29] Este modo de seleção de amostras permite criar diferentes possibilidades de conjuntos amostrais para compor cada árvore de decisão. Este artifício amplia o ranking da matriz de amostral, o que é importante quando se trabalha com uma grande quantidade de variáveis e poucas amostras.[30]

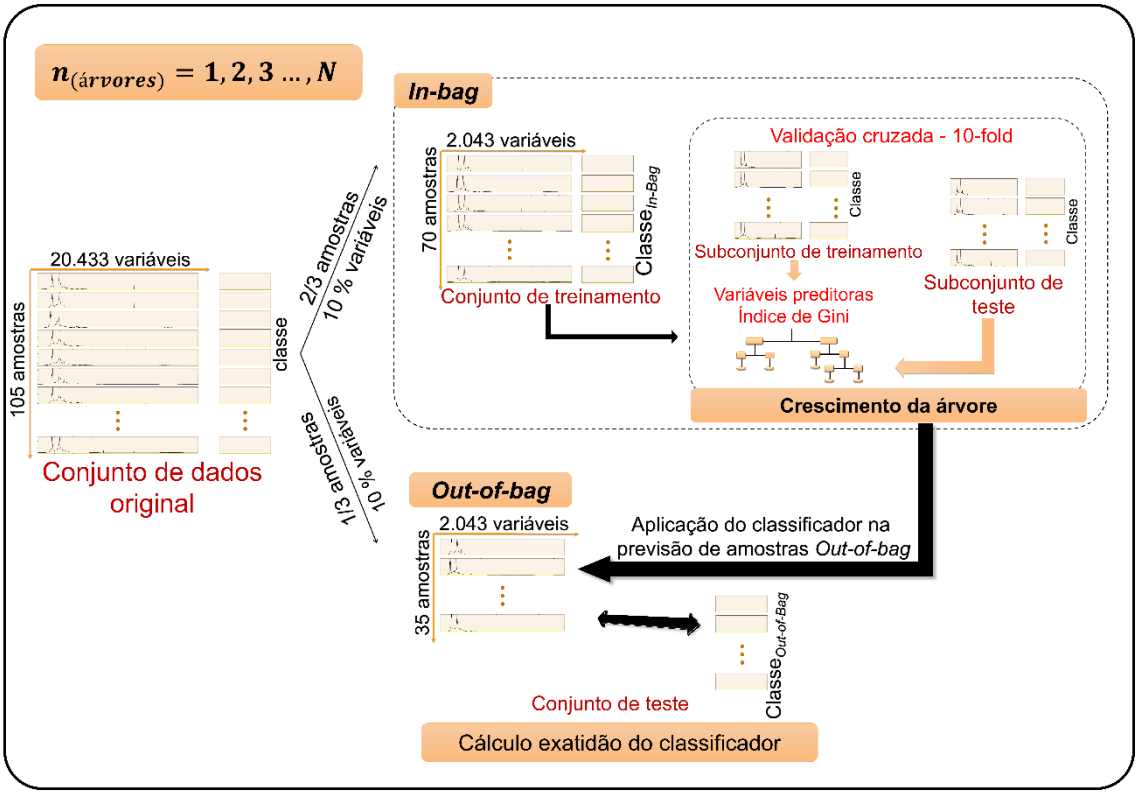

Figura 3.1 Construção do modelo de RF.

Durante a etapa in bag, o conjunto treinamento é dividido em subconjuntos de treinamento e teste utilizando o método de k-fold (com k=10) semelhante a uma validação cruzada. Concomitante à separação dos subconjuntos, variáveis preditoras são identificadas dentro da seleção randômica pelo critério de Gini e são utilizadas para compor os nós durante a construção da árvore de classificação.[31,32] Em seguida, é possível alimentar essa árvore com o conjunto teste (out of bag), que não participou da construção, a fim de classificá-las e verificar a exatidão (desempenho) do modelo criado para esse novo conjunto de dados.[31,32]

Todo esse processo de construção de uma única árvore é repetido iterativamente até que se conclua o número total de árvores (N).[12] Na literatura o número total de árvores preditoras é variável, há relatos de 500,[26,28] 1000,[33] e 2000[34] árvores. A predição com uso de vários classificadores independentes é vantajosa por ter distribuída a variância do conjunto de dados.[35,36] Cada classificador individual (CART) contribui com um voto final sobre a classe predita para o conjunto teste, essas decisões independentes são combinadas em um único classificador que resulta na classe com maior número de votos.[13,36]

3.3.3.3 Seleção de variáveis

Após a construção dos modelos de RF, foi avaliada a exatidão e selecionados os modelos com melhor desempenho (exatidão) para a seleção das variáveis de maior contribuição na discriminação das amostras. Um algoritmo interno foi acrescentado ao script de comandos, para a identificação das variáveis preditoras nas árvores de classificação com exatidão acima de 80%, dentro do modelo selecionado. Depois de identificadas as variáveis preditoras, foi realizada uma combinação dessas variáveis pelas suas respectivas regiões espectrais, nos espectros de RMN de ^{1}H e de ^{13}C, de acordo com as atribuições sugeridas por Poveda[37] (Tabela B.1, Tabela B.2 - Anexo B). Em seguida, calculamos a frequência relativa de todas as regiões em relação a que apresentou maior contribuição de variáveis. Por último, realizamos as atribuições aos compostos químicos presentes nessas variáveis,[37] para interpretação da contribuição química desses compostos na discriminação das amostras e consequente relação com o ponto de fluidez em petróleo.

3.4 Resultados e discussão

O perfil químico das 105 amostras de petróleo, representados por histogramas (Figura B.1, Anexo B), mostram uma distribuição de amostras em relação a algumas propriedades físico-químicas importantes na indústria de petróleo. A distribuição de amostras em relação à gravidade API (Figura B.1b) compreende uma ampla faixa (11,4 a 52,4), representativa dos petróleos brasileiros que varia de extra leves a asfálticos. [17,18,38] Entretanto, apresenta maior representatividade amostral na faixa entre 20 e 32, correspondentes aos óleos pesados e médios.[17,18,38] O histograma do teor de polares (resinas e asfaltenos) (Figura B.1e) apresentou uma variação de 0 a 43,4%, com uma parcela significante de amostras distribuídas na faixa de 13 a 30%. Petróleos com valores baixos para a gravidade API (óleos pesados) geralmente apresentam maior teor de polares[17], fato também sugerido pelos histogramas para esses dois parâmetros (Figuras B.1b e B.1e).

A distribuição pelo teor de saturados (Figura B.1c), apresenta uma ampla faixa variando de 36,7 a 88,0%, com uma maior expressividade amostral entre 47 e 62,4%. Pelo ponto de fluidez máximo (Figura B.1a), tem-se uma ampla distribuição de amostras, variando entre -54 ºC a 39 ºC. As faixas entre -35 a -26 ºC e 2 ºC a 20 ºC apresentam maior quantidade de amostras. O teor de aromáticos apresenta

distribuição entre os valores de 7 a 37,2%, com uma maior representatividade amostral entre os valores 19 e 28%.

Considerando o perfil amostral pelas propriedades demonstradas e avaliando a relação de ponto de fluidez a outras propriedades físico-químicas, a Figura 3.2 mostra a correlação entre ponto de fluidez máximo, teor de aromáticos (aromáticos e polares) e teor de saturados. Coeficientes de correlação calculados, para estas propriedades e API, assim como teste da significância estatística desses coeficientes, são mostrados no Anexo B (Tabelas B.1 e B.2).

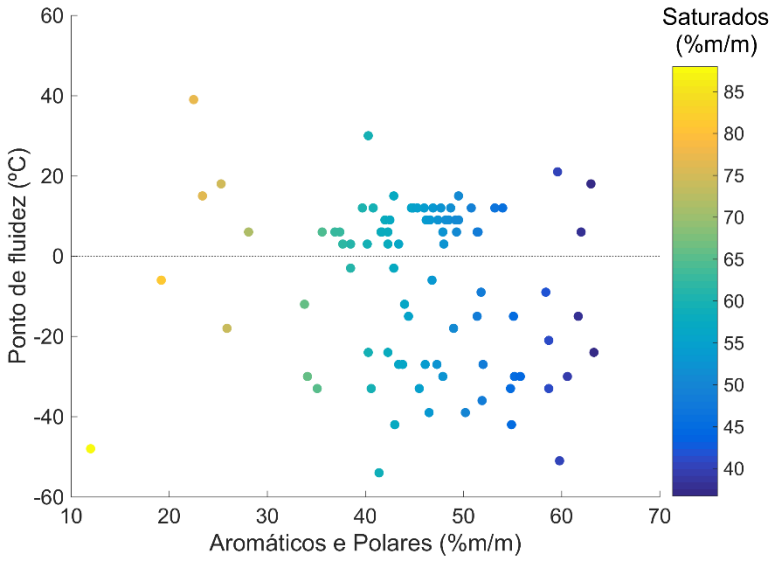

Figura 3.2 Gráfico de correlação entre ponto de fluidez, teor de aromáticos (aromáticos e polares) e teor de saturados.

O gráfico de correlação (Figura 3.2) relaciona três propriedades físico-químicas das amostras de petróleo. É possível observar um aumento gradativo no teor de aromáticos (12 a 63%) e uma redução do teor de saturados em proporções semelhantes. Ao traçar um limite em 0 ºC para o PFM são obtidos dois conjuntos amostrais. O conjunto superior contém amostras mais agrupadas e próximas a esse limite, em contrapartida, o conjunto inferior apresenta dispersão entre amostras e menor proximidade ao valor limite. Uma pequena parcela de amostras com porcentagens inferiores a 40% de aromáticos apresentam maiores teores de saturados e tendenciam a pontos de fluidez próximos ou superiores a 0 ºC. Em porcentagens entre 40 e 55% de aromáticos ocorre um equilíbrio com frações de saturados nos óleos. Nesta região do gráfico há uma maior concentração das

amostras. Existe também uma proporção similar de amostras com valor de PFM acima e abaixo do limite. De forma geral, o gráfico de correlação mostra que a ampla faixa para o PFM dos diferentes tipos de petróleos brasileiros em estudo está relacionada ao equilíbrio entre as frações de saturados e aromáticos. Na literatura, tem sido discutido em termos de compostos saturados[18,39] ou de aromáticos e polares. [17,18,40,41]

Os resultados obtidos pelo gráfico de correlação (Figura 3.2) também mostram a complexidade de se obter alguma relação uniforme e direta do PFM ao longo do gráfico com os teores de aromáticos e saturados. Nesse sentido, Sad et al. (2014)[18] reportam que estudos sobre o comportamento do PFM durante o processo de produção de diferentes tipos de óleo e sua correlação com outras propriedades de caracterização estão escassos na literatura por se tratar de uma propriedade independente das variáveis que influenciam nos demais parâmetros. A compreensão do PFM em petróleos é importante porque define as condições operacionais, permite calcular dimensionamento e permite ações preventivas e melhoria de processos, o que facilita a fluidez do óleo e reduz a incidência de incrustações em função da formação de cristais a partir de compostos parafínicos e precipitação de grupos poliaromáticos como os asfaltenos durante o transporte.[18,39] A precipitação de asfaltenos e parafinas representa uma séria preocupação para a indústria petrolífera, exigindo grandes investimentos de capital para remoção, prevenção da formação de depósitos e financiamento de pesquisa destinados a melhor esclarecer a composição química do petróleo, visto que, a estabilidade do óleo está relacionada ao equilíbrio dos teores de compostos aromáticos (incluindo resinas e asfaltenos), naftênicos e parafínicos.[40,42] Nessa perspectiva, dados espectroscópicos contendo informações sobre a caracterização do petróleo a nível molecular permitem encontrar relações espectrais com o PFM do óleo. Portanto, essas informações viabilizam discriminar amostras de petróleo quanto ao PFM e verificar se essa propriedade possui maior correlação com compostos saturados ou aromáticos.

Neste trabalho, os perfis químicos baseados em dados espectroscópicos das amostras de petróleo bruto podem ser observados pelo espectro de RMN de ^{1}H (Figura 3.3). Primeiramente, a variação dos deslocamentos químicos nos espectros foi corrigida com alinhamento de sinais pelo método de correlação de intervalos. É possível destacar o efeito do alinhamento, na região de deslocamento químico de 1,25 a 1,3 ppm, antes do alinhamento (Figura 3.3a) e após alinhamento (Figura 3.3b). O

mesmo procedimento foi realizado nos espectros de RMN de ^{13}C (Figura B.2, Anexo B).

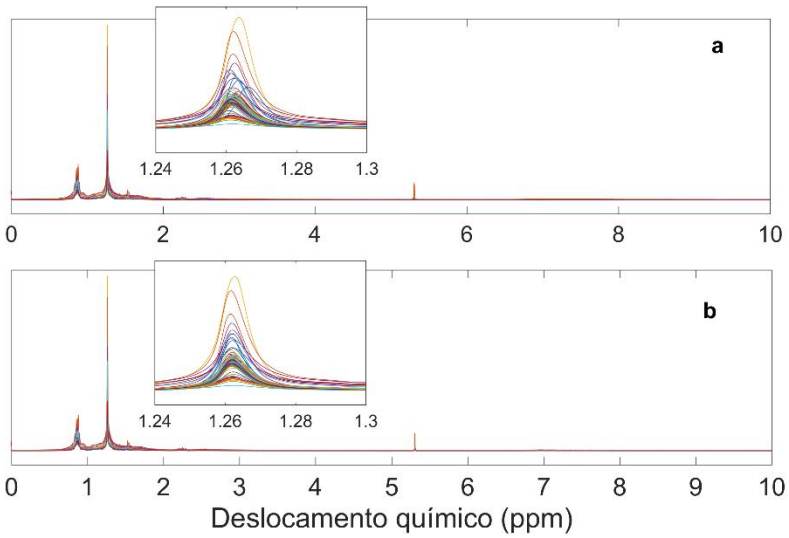

Figura 3.3 Espectros de RMN de ^{1}H das 105 amostras de petróleos utilizadas nos modelos de classificação. (a) dados originais e (b) após o alinhamento pelo algoritmo icoshift.

Nos espectros de RMN de ^{1}H regiões de interesse podem ser destacadas baseados nos assinalamentos estruturais de Poveda e Molina (Tabela C.1, Anexo C)[37]: 0,1-4,5 ppm representam os sinais relacionados a H do grupo de parafínicos e naftênicos associados ou não aos sistemas aromáticos, 6,0-9,0 ppm estão os H aromáticos e 9,0-10,0 ppm H mais polares de aldeídos e ácido carboxílicos. Analogamente, nos espectros de RMN de ^{13}C (Tabela C.2, Anexo C) podemos destacar as regiões: 3,0-85,0 ppm relacionados a C parafínicos e naftênicos, 85,0-160,0 ppm C aromáticos e 160,0-220,0 ppm C dos grupos polares de cetonas, quinolinas, aldeídos e ácidos.[37]

Os principais resultados da classificação por RF das amostras de petróleo a partir das informações espectrais estão apresentados na Tabela 3.1. Os modelos construídos com diferentes pré-processamentos mostraram diferentes desempenhos em classificar os dados espectrais.

Tabela 3.1 Resultados dos modelos de classificação para o PFM das 105 amostras de petróleo usando Floresta Randômica.

Modelo	Pré-processamento	Exatidão (%)	
		RMN de ^{1}H	RMN de ^{13}C
1	sem	78,0	71,0
2	SNV	90,2	70,3
3	autoescalamento	78,0	55,9
4	normalização	83,8	64,9
5	centrado	78,0	47,2
6	derivada	65,8	56,4

Os modelos de classificação obtidos por RMN de ^{1}H mostram melhor desempenho que os modelos de RMN de ^{13}C ao comparar os respectivos valores de exatidão para cada modelo. Dentre os modelos RMN de ^{1}H usando SNV como pré-processamento o modelo 2 apresentou melhor exatidão (90,2%). O modelo 1 de RMN de ^{13}C, sem pré-processamento mostrou melhor desempenho (71,0%) que os demais para esta técnica.

O modelo 2 (RMN de ^{1}H) foi selecionado para identificação das variáveis, de maior contribuição na discriminação das amostras. A partir dessa identificação foi realizado o cálculo da frequência dessas variáveis nos classificadores independentes com maior exatidão, desse modelo (Figura 3.4).

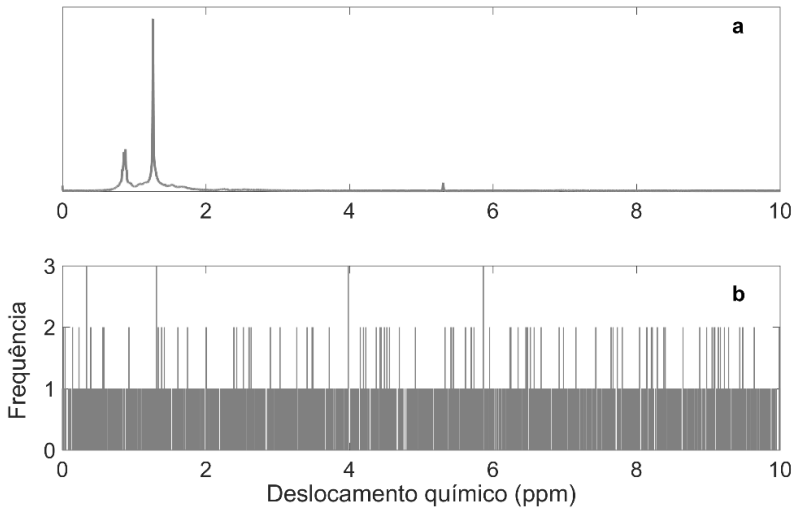

Figura 3.4 Identificação das variáveis preditoras; (a) deslocamento químico no espectro de RMN ^{1}H (b) gráfico da frequência no deslocamento químico.

Com as variáveis identificadas, é possível verificar as regiões espectrais assinaladas por Poveda e Molina[37] que mais influenciaram nos modelos de classificação (Fig. 3.5). No modelo 2 a região em que houve maior contribuição das variáveis corresponde aos H parafínicos e naftênicos, CH, CH_2 e CH_3, ligados a sistemas aromáticos na posição α (2,0-4,5 ppm). Em seguida, outra região de destaque é a de H aromáticos ligados a C aromáticos mono, di e poliaromáticos (7,2-9,0 ppm), além da contribuição de variáveis correspondentes a H aromático ligado a anéis monoaromáticos (6,0-7,2 ppm) e H de aldeídos e ácido carboxílicos (9,0–10,0 ppm), compreendendo a contribuição de H na fração de polares. Na região entre 4,5 e 6,0 ppm compreende a intensidade do sinal do solvente.

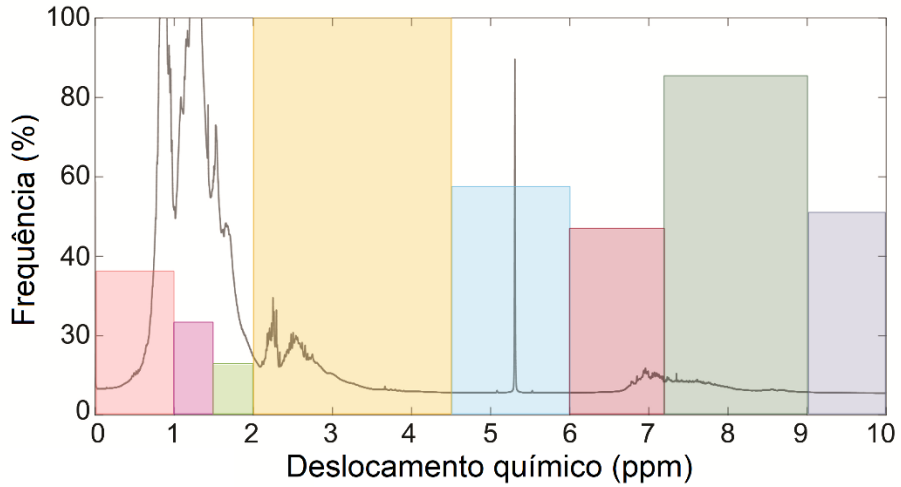

Figura 3.5 Espectro de RMN de ^{1}H e peso das variáveis (frequência relativa) por região espectral.

Para identificação das variáveis que mais contribuíram para a classificação com dados de RMN de ^{13}C, foi selecionado o Modelo 1. As duas regiões de maior peso das variáveis são atribuídas aos compostos polares (160 a 220 ppm) e C de sistemas benzonaftênicos e aromáticos (85 a 129 ppm) (Figura 3.6).

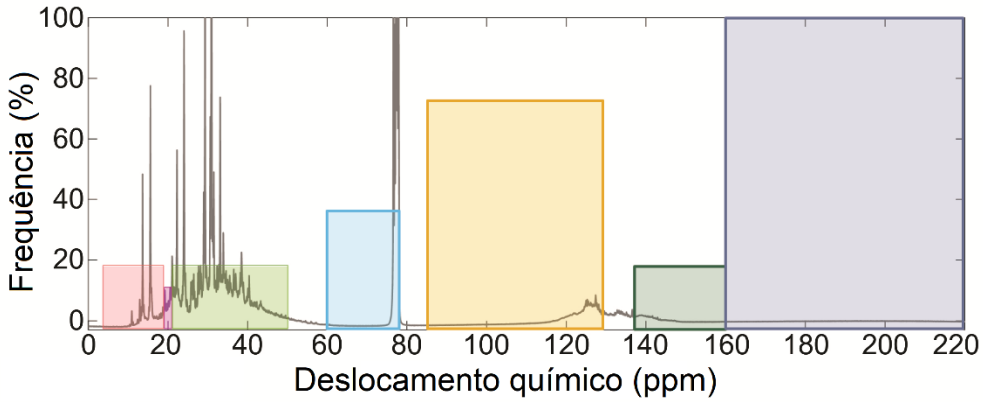

Figura 3.6 Espectro de RMN de ^{13}C e peso das variáveis (frequência relativa) por região espectral.

Pelo modelo de classificação de RF com dados de RMN de ^{13}C, identificamos uma menor contribuição das variáveis relacionadas aos compostos mono, di e poliaromáticos. Diferentemente dos modelos com dados de RMN de ^{1}H no qual se observou maior contribuição de H desses sistemas. O resultado da classificação pelos dados de RMN de ^{13}C sugere maior relação do ponto de fluidez com componentes polares, aromáticos e naftênicos entre sistemas aromáticos de poucos anéis, o que reforça a sugestão de uma interação de equilíbrio entre essas frações com o comportamento reológico do óleo.

Sad et al. (2014)[18] usaram um método não supervisionado e identificaram algumas correlações entre diferentes propriedades e ponto de fluidez. Seus resultados mostram uma correlação maior entre os asfaltenos e o ponto de fluidez, comportamento diferente que era esperado, que seria uma maior relação entre parafínicos e ponto de fluidez.[17,18,39,40,41]

Neste estudo, foi verificado não somente uma relação com aromáticos, mas também de compostos presentes em outras regiões espectrais. Essas divergências de resultados deste trabalho com os de Sad et al., podem ser explicadas pelas diferentes características de grupo amostral, como faixa de API (13,7 a 29,2) e de ponto de fluidez (de -30 °C a 6 ºC), mais estreitos do que neste trabalho. Isso reforça que o ponto de fluidez esteja relacionado a um estado de equilíbrio entre compostos parafínicos e aromáticos na composição do óleo, e não exclusivamente uma classe isolada de compostos.

Oliveira et al. (2018)[40] identificaram que ponto de fluidez sofreu influência de variações de temperaturas em amostras de petróleo e sugeriram uma relação de

elucidação do pondo de fluidez pela distinção de conteúdos de saturados e aromáticos entre as amostras. Neste trabalho, verificamos que há um equilíbrio entre as diferentes frações, destacando-se os grupos aromáticos e polares que contribuem para o comportamento de ponto de fluidez das amostras.

3.5 Conclusões

O método de floresta randômica é eficaz para discriminar amostras de óleo em relação ao ponto de fluidez máximo, obtendo resultados com precisão de 90% para dados de RMN de ^{1}H e 71% para dados de RMN de ^{13}C. A metodologia proposta, neste trabalho, para selecionar as variáveis mais frequentes na construção das árvores com maior exatidão, é eficiente na identificação das variáveis que melhor contribuem para os modelos de classificação. A partir do método de seleção de variáveis aplicado, é possível verificar uma relação de equilíbrio entre o ponto de fluidez e outras propriedades físico-químicas do óleo que influenciam no comportamento reológico do óleo.

3.6 Referências Bibliográficas

1. LAVINE, B. K. Chemometrics. *Analytical Chemistry*, **72**, 91-97, 2000.

2. LU, D.; WENG, Q. A survey of image classification methods and techniques for improving classification performance. *Int. J. Remote Sens*., **28**, 823–870, 2007.

3. SAYAGO, A.; BELTRÁN, R.; GONZÁLES-DOMÍNGUES, A. Combination of complementary data mining methods for geographical characterization of extra virgin olive oils based on mineral composition. *Food Chemistry*, **261**, 42–50, 2018.

4. BELGIU, M.; DRAGUT, L. Random forest in remote sensing: A review of applications and future directions, *ISPRS J. Photogramm Remote Sens*., **114**, 24–31, 2016.

5. BREIMAN, L. Random Forests. *Machine Learning*, **45**, 5-32, 2001.

6. LIU, W.; LIU, C.; HU, X.; YANG, J.; ZHENG, L. Application of terahertz spectroscopy imaging for discrimination of transgenic rice seeds with chemometrics, *Food Chemistry*. **210**, 415–421, 2016.

7. LEE, S.; CHOI, H.; CHA, K.; CHUNG, H. Random forest as a potential multivariate method for near-infrared (NIR) spectroscopic analysis of complex mixture samples: Gasoline and naphtha. *Microchemical Journal*, **110**, 739–748, 2013.

8. STROZIER, E. D.; MOONEY, D. D.; FRIEDENBERG, D. A.; KLUPINSKI, T. P.; TRIPLETT, C. A. Use of Comprehensive Two-Dimensional Gas Chromatography with Time-of-Flight Mass Spectrometric Detection and Random Forest Pattern Recognition Techniques for Classifying Chemical Threat Agents and Detecting Chemical Attribution Signatures, *Analytical Chemistry*, **88**, 7068−7075, 2016.

9. SHANG, W.; HUANG, H.; ZHU, H.; LIN, Y.; QU, Y.; WANG, Z. A novel feature selection algorithm for text categorization, *Expert Systems with Applications*. **33**, 2007, 1–5, 2007.

10. BASHIR, S.; QAMAR, U.; KHAN, F. H.; JAVED, M. Y. MV5: A Clinical Decision Support Framework for Heart Disease Prediction Using Majority Vote Based Classifier Ensemble, *Arabian Journal for Science and Engineering*, **39**, 7771–7783, 2014.

11. PARK, H.; KWON, H. C. Improved Gini-Index Algorithm to Corrct feature-selection bias in text classification, *IEICE Trans. Inf. & Syst*, E94-D, 855-865,2011.

12. CAO, D. S.; HU, Q. N.; XU, Q. S.; YANG, Y. N.; ZHAO, J. C.; LU, H. M.; ZHANG, L. X.; LIANG, Y. Z. In silico classification of human maximum recommended daily dose based on modified random forest and substructure fingerprint. *Analytica Chimica Acta*, **692**, 50–56, 2011.

13. ZHANG, T.; XIA, D.; TANG, H.; YANG, X.; LI, H. Classification of steel samples by laser-induced breakdown spectroscopy and random forest, *Chemometrics and Intelligent Laboratory Systems*, **157**, 196-201, 2016.

14. ISMAIL, S. N.; MAULIDIANI, M.; AKHTAR, M. T.; ABAS, F.; ISMAIL, I. S.; KHATIB, A.; ALI, N. A. M.; SHAARI, K. Discriminative Analysis of Different Grades of Gaharu (Aquilaria malaccensis Lamk.) by 1H-NMR-Based Metabolomics Using PLS-DA and Random Forests Classification Models, *Molecules*, **22**, 1612-1625, 2017.

15. CANIZO, B. V.; ESCUDERO, L. B.; PÉREZ, M. B.; PELLERANO, R. G. Intra-regional classification of grape seeds produced in Mendoza province (Argentina) by multi-elemental analysis and chemometrics tools. *Food Chemistry*, **242**, 272–278, 2018.

16. ASTM, D5853 “Standard test method for pour point of crude oils. WestConshohocken PA: ASTM International”, 2011.

17. FILGUEIRAS, P.R.; PORTELA, N.A.; SILVA, S. R. C.; CASTRO, E. V. R.; OLIVEIRA, L. M. S. L.; DIAS, J. C. M.; NETO, A. C.; ROMÃO, W; POPPI, R. J. Determination of saturates, aromatics, and polars in crude oil by ^{13}C NMR and support vector regression with variable selection by genetic algorithm. *Energy Fuels*. **30,**1972-1978, 2016.

18. SAD, C. M. S.; LACERDA, J. R. V.; FILGUEIRAS, P. R.; RIGONI, V. S.; BASSANE, J. F. P.; CASTRO, E. V. R.; PEREIRA, K. S.; SANTOS, M. F. P. Limitations of the pour point measurement and the influence of the oil composition on its detection using principal component analysis, *Energy Fuels*, **28**, 1686–1691, 2014.

19. SAVORANI, F.; TOMASI, G.; ENGELSEN, S. B. Icoshift: A versatile tool for the rapid alignment of ^{1}D NM spectra, *J. Magn. Resson.,* **202**, 190-202, 2010.

20.TOMASI, G.; SAVORANI, F.; ENGELSEN, S. Icoshifit: an effective tool for the alignment of chromatographic data. *J. Chromatogr.,* **1218**, 7832-7840, 2011.

21. BLANCO, M.; COELHO, J.; ITURRIAGA, H.; MASPOCH, S.; PEZUELA, C. Effect of data preprocessing methods in near-infrared diffuse reflectance spectroscopy for the determination of the active compound in a pharmaceutical preparation. *Appl. Spectrosc.*, **51**, 240-246, 1997.

22. GELADI, P.; KOWALSKI, B. R. Partial least-squares regression: a tutorial. *Anal. Chim. Acta*, **185**, 1-17, 1986.

23. BARNES, R. J.; DHANOA, M.S.; LISTER, S. J. Standard Normal Variate transformation and De-trending of near-infrared diffuse reflectance spectra. *Appl. Spectrosc*. 43, 772-777, 1989.

24. FEARN, T.; RICCIOLI, C.; GARRIDO-VARO, A.; GUERRERO-GINEL, J. E. On the geometry of SNV and MSC, *Chemometrics intel. Lab. Syst.* **96**, 22-26, 2009.

25. SVETNIK, V. LIAW, A.; TONG, C.; CULBERSON, J. C.; SHERIDAN, R. P.; FEUSTON, B. P. Random forest: a classification and regression tool for compound classification and QSAR modeling. *J. Chem. Inf. Comput. Sci*. **43**, 1947–1958, 2003.

26. POLIANCHIK, D. E.; GRIGORO'EV, V. Y.; SANDAKOV, G. I.; YARKOV, A. V.; BACHURIN, S. O.; RAEVSKII, O. A. Binary Cassification of CNS and PNS drugs. *Pharmaceutical Chemistry Journal,* **50**, 801-804, 2016.

27. SANTANA, F. B.; SOUZA, A. M.; POPPI, R.J. Visible and near infrared spectroscopy coupled to random forest to quantify some soil quality parameters, *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, **191**, 454–462, 2018.

28. DONG, J.; GUO, W.; LIU, D. Discrimintion of "Hayward" Kiwifruits Treated with Forchlorfenuron at different concentrations using hyperspectral imaging technology. *Food Analytical Methods*,**19**, 477-486, 2017.

29. BIAU, G. Analysis of a Random Forests Model. *Journal of Machine Learning Research*, **13**, 1063-1095, 2012.

30. ZHENG, L.; WATSON, D. G.; JOHNSTON, B. F.; CLARK, R. L.; EDRADA-EBEL, R.; ELSEHERI, W. A chemometric study of chromatograms of tea extracts by correlation optimization warping in conjunction with PCA, support vector machines and random forest data modeling, *Analytica Chimica Acta*, **642**, 257–265, 2009.

31. AFANADOR, N. L.; SMOLINSKA, A.; TRAN, T. N.; BLANCHET, L. Unsupervised random forest: a tutorial with case studies, *Journal of Chemometrics*. **30**, 232–241, 2016.

32. CHO, J. H.; KURUP, P.U. Decision tree approach for classification and dimensionality reduction of electronic nose data. *Sensors and Actuators B: Chemical*, **160**, 542-548, 2011

33. VIGNEAUA, E.; COURCOUX, P.; SYMONEAUX, R.; GUÉRIN, L.; VILLIÈRE, A. Random forests: A machine learning methodology to highlight the volatile organic

compounds involved in olfactory perception, *Food Quality and Preference*, **68**,135–145, 2018.

34. LI, Y.; ZHANG, J. Y.; WANG, Y. Z. FT-MIR and NIR spectral data fusion: a synergetic strategy for the geographical traceability of Panax notoginseng. *Analytical and Bioanalytical Chemistry*, **410**, 91–103, 2018.

35. BREIMAN, L. Bagging predictors. *Machine Learning*, **24**, 123-140, 1996.

36. GISLASON, P. O.; BENEDIKTSSON, J. A.; SVEINSSON, J. R. Random Forests for land cover classification, *Pattern Recognition Letters*, **27**,294–300, 2006.

37. POVEDA, J. C.; MOLINA, D. R. Average molecular parameters of heavy crude oils and their fractions using NMR spectroscopy. *Journal of Petroleum Science and Engineering*, **84**, 1–7, 2012.

38. DUARTE, L. M.; FILGUEIRAS, P. R.; DIAS, J. C. M.; OLIVEIRA, L. M. S. L.; CASTRO, E. V. R.; OLIVEIRA, M. A. L. Study of Distillation Temperature Curves from Brazilian Crude Oil by 1 H Nuclear Magnetic Resonance Spectroscopy in Association with Partial Least Squares Regression. *Energy Fuels*, **31**, 3892-3897, 2017.

39. RODRIGUES, E. V. A.; SILVA, S. R. C.; ROMÃO, W.; CASTRO, E. V. R.; FILGUEIRAS, P. R. Determination of crude oil physicochemical properties by high-temperature gas chromatography associated with multivariate calibration. *Fuel*, **220**, 389–395, 2018.

40. OLIVEIRA, L. M. S. L.; NUNES, R. C. P.; RIBEIRO, Y. L. L.; COUTINHO, D. M.; AZEVEDO, D. A.; DIAS, J. C. M.; LUCAS, E. F. Wax Behavior in Crude Oils by Pour Point Analyses. *J. Braz. Chem. Soc*. 29, 2158-2168, 2018.

41. SANTOS, F.D.; SAD, C. M. S.; BASSANE, J. F. P.; NETO, D. M. C.; FILGUEIRAS, P. R. SILVA, M.; SILVA, S. R. C.; CASTRO, E. V. R.; ROMÃO, W.; PEREIRA, L. B.; LACERDA, V. Improvement on Pour Point of Heavy Oils by Adding Organic Solvents, Rev. Virtual Quím., **9**, 2404-2413, 2017.

42. F.C. TOZZI, C.M.S. SAD, J.F.P. BASSANE, F.D. SANTOS, M. SILVA, P.R. FILGUEIRAS, H.P. DIAS, W. ROMÃO, E.V.R. DE CASTRO, V. L. Jr. Improving the physicochemical properties of Brazilian onshore and offshore crude oils using the production of blends. *Fuel*, **159**, 607-613, 2015.

4. APLICAÇÃO 2: DIFERENTES ESTRATÉGIAS PARA O USO DE FLORESTA RANDÔMICA EM DADOS DE RESSONÂNCIA MAGNÉTICA NUCLEAR

4.1 Introdução

Floresta randômica (RF) é um método de aprendizado de máquina reconhecido pela eficiência na classificação supervisionada de dados lineares e não lineares, especialmente, nas áreas de análise genética, metabolômica, análise de imagens médicas, controle de qualidade de alimentos e petróleo.[1-4] A potencialidade da técnica de RF em classificação é confirmada quando associada à métodos analíticos que permitem aquisição de grande quantidade de informações como: cromatografia, espectrometria de massas, espectroscopia de infravermelho, ressonância magnética nuclear de hidrogênio (RMN de ^{1}H) e carbono (RMN de ^{13}C).[1-4]

A técnica de RMN, em especial, fornece informações detalhadas sobre grupos funcionais e caracterização da amostra a nível molecular e a quantidade de variáveis em estudo pode exceder de 65 mil, a depender da faixa espectral analisada. Em contrapartida, aplicações de RF para solucionar problemas de classificação com grande quantidade de dados e conjuntos amostrais mais extensos pode gerar um elevado custo computacional, além de demandar maior tempo.[5-7] Nesse sentido, para obter modelos mais rápidos e parcimoniosos torna-se importante o uso de métodos que selecionem as variáveis de maior relevância, anteriormente a aplicação de RF. Esses métodos são capazes de proporcionar uma redução da dimensionalidade de dados e podem destacar as variações que diferenciam as informações entre os espectros.[7]

A técnica de RF foi criada em 2001, pela fusão dos algoritmos CART (*Classification and Regression Trees*) e *bootstrapping aggregation* (*bagging*),[8] ainda há poucos relatos sobre as estratégias de seleção de variáveis em dados espectroscópicos combinadas a RF ou quando associada a análise de componentes principais (PCA).[9,10] Shen et al. (2018),[9] utilizam essa metodologia para discriminar alfaces contaminadas com diferentes concentrações de cádmio (Cd) a partir da espectroscopia de emissão em plasma induzido por laser (LIBS). Os autores

classificaram as informações por duas formas: a primeira usando somente RF, e a segunda, associando a PCA ao RF. O resultado obtido apenas com RF foi de 80% de exatidão, considerado baixo quando comparado com a fusão da PCA e RF. A segunda forma mostrou 100% de exatidão na classificação das amostras de alface. Essa mesma ideia, foi aplicada por Amjad et al. (2018)[10] para discriminar amostras de leite pela origem biológica (búfala, vaca, cabra e humano). Os espectros de Raman dos tipos de leite demonstraram diferentes intensidades decorrentes das variações quanto ao teor de proteínas, lactose, gordura e outros nutrientes. Foram aplicados os métodos combinados de PCA e RF nos dados e os resultados alcançaram valores de 93,7% de exatidão para classificar o leite nas quatro classes.

Apesar de poucos estudos relacionados a combinação de técnicas de seleção de variáveis com RF, é possível verificar a eficiência do método no processo de classificação de matrizes com diferentes níveis de complexidade. Não há abordagem na literatura de estudos voltados sobre a associação de RF a outros métodos para redução de dimensionalidade, identificação de variáveis utilizadas pelo classificador, assim como aplicações deste modo em dados de RMN.

4.2 Objetivo

O objetivo deste trabalho é comparar a eficiência (exatidão) de RF quando associado ou não aos métodos de seleção de variáveis por PCA e discriminante de Fisher (FD – *Fisher's Driscriminant*) para distinguir as amostras de petróleo em relação ao Número de Acidez Total (NAT), utilizando RMN de ^{1}H e ^{13}C. Para os modelos RF obtidos com toda faixa espectral e FD, esse trabalho também pretende identificar e comparar as variáveis mais importantes utilizadas durante o processo de construção das árvores.

4.3 Metodologia

Neste trabalho, foram utilizadas 126 amostras de óleo bruto, variando de leves a pesados (faixa de gravidade API de 11,4 a 54,0).[11] O NAT foi determinado de acordo com o método padrão ASTM D664[12] apresentando variação de 0,03 a 4,96 mg $KOH \cdot g^{-1}$ (Figura 4.1). Foram adquiridos espectros de RMN de ^{1}H (20.433 variáveis) e ^{13}C (61.606 variáveis). Os valores de NAT foram usados para avaliar as suas possíveis

correlações com os sinais espectrais dos compostos do petróleo. Previamente, foi verificado em todo o conjunto amostral o teor de água abaixo de 0,5% (v/v).[13]

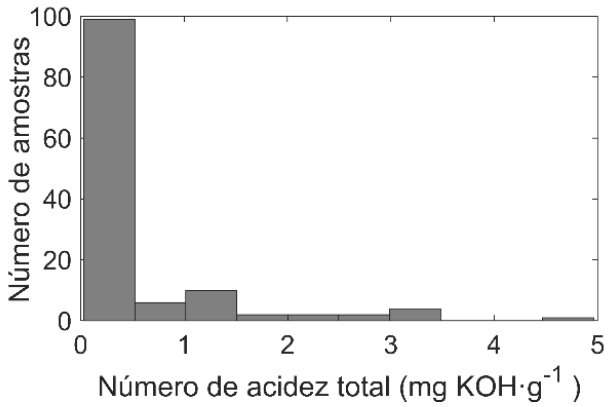

Figura 4.1 Histograma do conjunto amostral em relação ao número de acidez total.

Primeiramente, as amostras foram classificadas com base nos resultados de NAT, em que foi determinada uma concentração como valor referência (0,3 mg $KOH \cdot g^{-1}$). Dessa forma, cada classe teria um número representativo e proporcional de amostras. Assim, “classe 1” foi constituída por 78 amostras com valores de NAT abaixo ou igual a 0,3 mg $KOH \cdot g^{-1}$ e para a “classe 2” foram atribuídas as 48 amostras com valores de NAT acima de 0,3 mg $KOH \cdot g^{-1}$. Na indústria do petróleo adota-se valores de número de acidez total (NAT) abaixo de 0,5 mg $KOH.g^{-1}$ para óleos não acidos e acima de 0,5 mg $KOH.g^{-1}$ para óleos ácidos. Entretanto, é obervado que alguns óleos com NAT abaixo de 0,5 mg $KOH.g^{-1}$ possui característica corrosiva isso se deve a natureza das espécies ácidas presentes no óleo. A propriedade NAT possui valor de reprodutibilidade elevado (o valor depende do NAT total medido) pelo método ASTM D664,[12] assim, para garantir baixa acidez e baixo poder corrosivo do óleo foi adotado o valor de 0,3 mg $KOH.g^{-1}$ como limite, para manter uma forma mais conservadora na determinação de um óleo com baixa acidez e poder corrosivo. A partir dessas classes, desenvolvemos os modelos de classificação supervisionada por RF (Figura 4.2), em dois modos: 1) utilizando toda a faixa espectral (primeiro modo, Modelo RF); 2) selecionando variáveis por PCA ou FD, anteriormente a classificação por RF (segundo modo, Modelo PCA-RF ou Modelo FD-RF). Em seguida, os resultados dos modos de obtenção dos modelos RF foram avaliados e comparados com base na exatidão. Para os dois modos, alguns modelos foram construídos com aplicação de métodos de pré-processamento, que também foram avaliados segundo

o desempenho. Para identificar as variáveis com maior contribuição na classificação das amostras, foram selecionados os modelos com melhor exatidão. Para isso, foram utilizados o Modelo RF (do primeiro modo) e o Modelo FD-RF do segundo modo.[4]

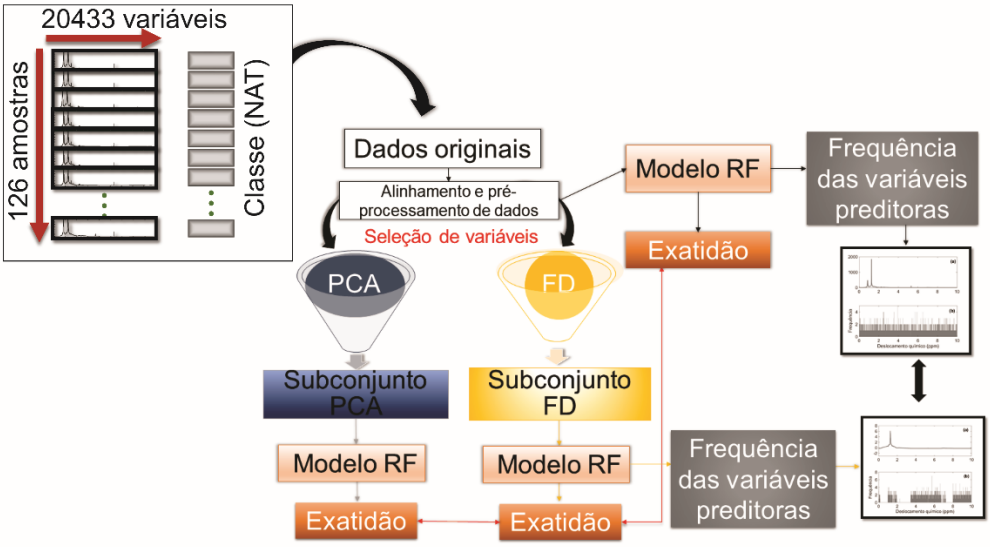

Figura 4.2 Etapas da construção dos modelos de RF. Usando toda a faixa espectral e com métodos de seleção de variáveis para os dados de RMN.

4.3.1 Aquisição dos espectros do RMN de ^{13}C

Os espectros de RMN de ^{13}C foram obtidos a 27 ºC, utilizando um instrumento Varian VNMRS-400 com um campo magnético de 9,4 T, 10 mm de Banda Larga ^{15}N-^{31}P (^{1}H-^{19}F). As amostras foram preparadas dissolvendo-se uma massa de óleo (correspondente a 40% do volume final da solução em clorofórmio deuterado, $CDCl_3$) contendo 0.05 mol·L^{-1} de acetilacetonato de cromo III, 3 mL de volume total. As condições instrumentais estabelecidas foram: frequência, 100,51 MHz para o núcleo ^{13}C; janela espectral, 25510,2 Hz; tempo de aquisição, 1,285 s; tempo de espera, 7 s; pulso, 90° (14,2 µs); número de transientes, 1000; e modo de desacoplamento, nny. O desacoplamento foi desligado durante o pulso e também no tempo de espera, e novamente foi conectado na etapa de aquisição de dados.

4.3.2 Aquisição dos espectros do RMN de ^{1}H

Os espectros de RMN de ^{1}H foram obtidos a 27 ºC, utilizando um instrumento Varian VNMRS-400 com um campo magnético de 9,4 T, 5 mm de Banda Larga ^{1}H-

19F. As amostras foram preparadas dissolvendo uma massa de óleo (correspondendo a 5% do volume final da solução em diclorometano, CH_2Cl_2), 600 µL do volume total. As condições instrumentais estabelecidas foram: frequência, 399,73 MHz; janela espectral, 6410,3 Hz; tempo de aquisição, 2,556 s; tempo de espera, 1 s; pulso, 90° (11.7 µs); número de transientes, 64.

4.3.3 Medidas do número de acidez total

O número de acidez total foi aferido experimentalmente por titulação potenciométrica, conforme o método padrão ASTM D664 (2011).[12] Nesse procedimento, uma alíquota do óleo cru foi dissolvida em uma solução com constituição volumétrica (%v/v) de 50% de tolueno PA, 49,5% de propan-2-ol PA e 0,5% de água ultrapura. Em seguida, essa mistura foi titulada com um padrão alcoólico de hidróxido de potássio (KOH) com concentração de 0,1 $mol \cdot L^{-1}$, para a determinação do número total de compostos ácidos presentes na amostra, dado em mg $KOH \cdot g^{-1}$.

4.3.4 Quimiometria

4.3.4.1 Pré-processamento e processamento de dados

Primeiramente, os espectros foram alinhados com o algoritmo icoshift para corrigir pequenos deslocamentos nos desvios químicos.[14,15] Esse alinhamento reduz a variação de informações não relacionadas à composição química entre as amostras. Após o alinhamento dos espectros, os dados foram processados. Os métodos de pré-processamento dos dados utilizados para avaliação foram: auto-escalamento, Variação Normal Padrão (SNV - *Standard Normal Variate*), centralização de média e derivada de primeira ordem. [16-19] A etapa de pré-processamento contribui significativamente para a análise quimiométrica, promovendo a atenuação dos efeitos causados pelos interferentes e pelas variações espectrais indesejadas. Após o pré-tratamento dos dados espectrais, geralmente, obtemos modelos mais robustos, precisos e parcimoniosos.[20] Vale ressaltar que toda a análise dos dados foi realizada com o software MATLAB 7.12.

4.3.4.2 Seleção de variáveis por Análise de Componentes Principais

O método de análise de componentes principais foi aplicado para a redução da dimensionalidade dados originais em componentes principais pela maximização

da variância dos dados. Neste trabalho foram utilizadas as 20 primeiras componentes principais, que acumularam uma variância superior a 99,9%. Por esse processo foram obtidos os escores contendo a informação de maior relevância dos espectros sobre as amostras. Dessa maneira, a matriz de scores foi utilizada como fator de entrada para os modelos de classificação com RF.

4.3.4.3 Seleção de variáveis por Discriminante de Fisher

A análise FD foi utilizada para selecionar variáveis que continham diferenças significativas entre as duas classes definidas com relação aos valores de NAT. A razão de Fisher (Equações 4.1, 4.2 e 4.3) é estabelecida como a variação classe-a-classe da variável independente dividida pela soma das variações dentro da própria classe.[21,22] Primeiramente foram calculadas as variações entre as classes (D).

$$D = \frac{\sum(\bar{x}_l - \bar{x})^2 ni}{(L-1)} \quad , \tag{4.1}$$

em que, n_i o número de amostras na enésima classe, $\bar{x}_l$ é a média da enésima classe, $\bar{x}$ a média global e L o número de classes. Em seguida é calculada a variação intra classe (d) (Equação 4.2).

$$d = \frac{\sum(\sum(x_{ij} - \bar{x})^2) - (\sum((\bar{x}_l - \bar{x})^2 ni))}{(N-L)} \tag{4.2}$$

em que x_{ij} é a enésima medida da enésima classe e N é o número total de amostras. Por fim é calculada a razão discriminante de Fisher (FD) (Equação 4.3).

$$FD = \frac{D}{d} \tag{4.3}$$

Se uma variável possuir um valor pequeno para a FD, significa que a variação dentro da sua classe (d) para aquela variável é maior do que entre as diferentes classes (D). Isso indica que essa determinada variável não será um fator importante para o processo de classificação. Por outro lado, se a variável em questão apresentar

um alto valor de FD, significa que a variação dentro da própria classe (d) é menor que a variação entre duas classes.

Com a aplicação de FD nos dados originais foram obtidas as variáveis que mais destacaram-se (maior peso) durante a discriminação de um espectro do outro. Em seguida, utilizou-se essas informações espectrais selecionadas para o desenvolvimento dos modelos RF. Após a redução por FD de aproximadamente 37% das variáveis de ^{1}H e 83,5% variáveis de carbono, os modelos RF foram construídos utilizando como dados de entrada as 12.807 variáveis de RMN de ^{1}H e as 9.900 variáveis provenientes do RMN de ^{13}C.

4.3.4.4 Modelo de classificação de Floresta Randômica

A partir do conjunto de dados de RMN de ^{1}H (variáveis 20.433) e ^{13}C (61.606 variáveis) das 126 amostras de petróleo, desenvolvemos os modelos de classificação por RF.[4] Todos os modelos foram obtidos com a seleção de amostras por bootstrap para formar o conjunto treinamento, chamado in bag (aproximadamente 2/3 do total), e conjunto teste, chamado out-of-bag (aproximadamente 1/3 do total). Em todos os modos de desenvolvimento dos modelos, o tempo e o custo de processamento da máquina foram comparados (Tabela D.1, Anexo D). Os modelos construídos utilizando toda a faixa espectral foram otimizados com 10% de variáveis para serem selecionadas randomicamente durante a construção das árvores. Nesse primeiro modo (Modelo RF), foram utilizadas 1.000 árvores para os sinais de RMN de ^{1}H e 2.000 com os dados de RMN de ^{13}C. Os modelos RF desenvolvidos no segundo modo, utilizando a PCA para a seleção de variáveis (Modelo PCA-RF) foram otimizados com 50% de variáveis para a seleção randômica. Esses modelos foram desenvolvidos com 500 árvores para os modelos baseados no RMN de ^{1}H e 1.500 árvores para o ^{13}C. No segundo modo, os modelos RF obtidos com a redução dos dados por FD, foram otimizados com 10% de variáveis para seleção randômica, 1.000 árvores para os modelos com RMN de ^{1}H e 1.500 árvores para os dados de RMN de ^{13}C.

4.3.4.5 Identificação de variáveis preditoras

Os modelos RF foram avaliados em relação ao nível de exatidão, utilizando os resultados obtidos para o primeiro modo (Modelo RF) e segundo modo com a redução de variáveis por FD (Modelo FD-RF). Em cada tipo de construção foi escolhido o

modelo que apresentou o melhor desempenho (exatidão), para a verificação das variáveis com maior contribuição na discriminação das amostras.[4] A identificação das variáveis preditoras foi realizada com as árvores com valores de exatidão acima de 70%, devido à presença constante dessas variáveis na construção dos melhores modelos. Por último, foi realizada uma combinação das variáveis preditoras identificadas com as respectivas regiões espectrais de acordo com as atribuições sugeridas por Poveda e Molina (2012)[23] (Tabela C.1 e Tabela C.2, Anexo C). Então foram calculadas as frequências relativas para todas as regiões do RMN de ^{1}H e ^{13}C foram realizadas usando a metodologia proposta por Lovatti et al. (2019)[4], tendo como referência a região que apresentou a maior contribuição (peso) como referência. Finalmente, os resultados foram interpretados conforme as contribuições químicas dos compostos presentes nessas variáveis, que possibilitaram explicar a discriminação das amostras e a consequente relação com o NAT em petróleo.

4.4 Resultados e discussão

A composição do petróleo bruto é predominantemente por hidrocarbonetos alifáticos, naftênicos e aromáticos. Esses grupos moleculares quando associados ao oxigênio formam compostos oxigenados, dentre eles o grupo funcional de ácidos carboxílicos em que estão incluídos os ácidos naftênicos. Esses ácidos podem ser representados pela fórmula $R(CH_2)_nCOOH$, onde R indica uma estrutura aromática ou ciclo-alifática. Nesse sentido, as técnicas de RMN de ^{1}H e ^{13}C podem ser adequadas para caracterização de petróleo a nível molecular, permitindo relacionar os dados espectrais com a acidez do óleo. Portanto, com estas informações é possível verificar se existe a ocorrência de discriminação entre as amostras e se os ácidos carboxílicos contidos nelas estão mais associados com os grupos aromáticos ou saturados (cadeias aberta e fechada). Estudos de classificação do petróleo quanto a acidez são de grande importância devido ao efeito corrosivo associado a essa classe de compostos. A produção mundial de petróleos ácidos aumentou e a matriz brasileira está listada entre as mais ácidas do mundo. Aumentar essa produtividade implica no desenvolvimento de novas técnicas para a compreensão e identificação de compostos de petróleo, especialmente ácidos naftênicos.[24,25]

Baseado nos assinalamentos estruturais de Poveda e Molina (2012)[23] (Tabela C.1, Anexo C), mostra que as regiões de interesse são: 9-10 ppm os sinais

relacionados ao H de aldeído e ácido carboxílico, 6-9 ppm para o H de aromáticos, 0,1-4,5 ppm para o H dos grupos alquílicos e naftênicos. Analogamente, nos espectros de RMN de ^{13}C (Tabela C.2, Anexo C) podemos destacar a região de 160-220 ppm referente ao sinal do C de ácidos carboxílicos, quinolinas, cetonas e aldeídos, a faixa de 85-160 ppm para o C dos aromáticos e os sinais de 3-85 ppm para o C dos grupos parafínicos e naftênicos (Figura D.1, Anexo D). Antes da construção dos modelos de classificação, os espectros de RMN foram alinhados com o algoritmo icoshift. A Figura 4.3 destaca o efeito dessa aplicação na região do deslocamento químico de 1,25 a 1,3 ppm, evidenciando o espectro antes do alinhamento (Figura 4.3a) e após (Figura 4.3b). O mesmo procedimento foi realizado nos espectros de RMN de ^{13}C (Figura D.1, Anexo D).

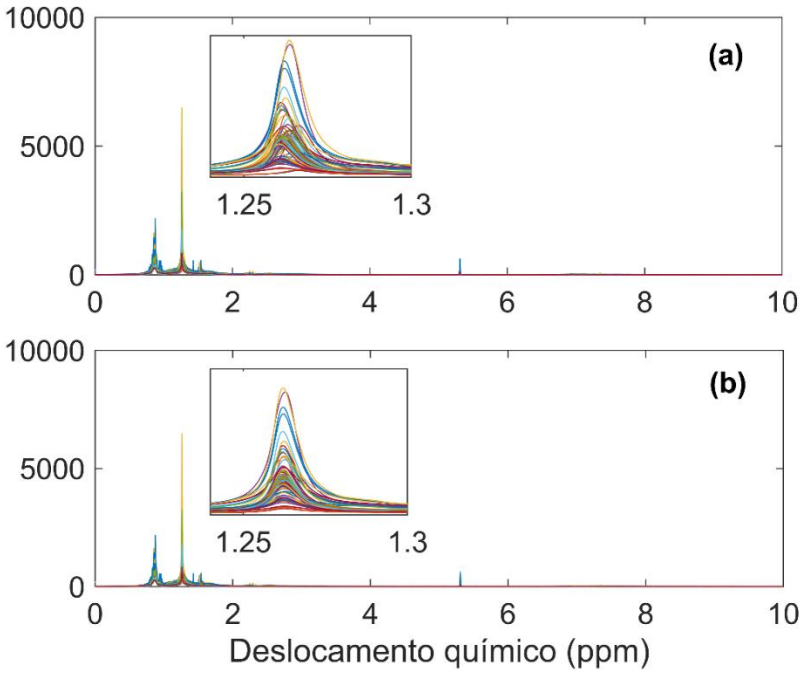

Figura 4.3 Espectros de RMN de ^{1}H das 126 amostras de petróleos utilizadas nos modelos de classificação. (a) dados originais e (b) após o alinhamento pelo algoritmo icoshift.

Os resultados (Tabela 4.1) de desempenho com os dados espectrais associados à classificação por RF nos dois modos (toda a faixa espectral e métodos de seleção de variáveis), mostraram que os melhores níveis de exatidão foram para aqueles os modelos com algum método seletor de variável. Para o RMN de ^{1}H, em comparação com as demais observações, a combinação de FD e RF alcançou o valor de exatidão superior (87,7%) quando empregou o pré-processamento SNV nas

12.807 variáveis selecionadas (Modelo FD-RF 2). O segundo melhor resultado (81,6%) foi construído com os dados de toda a faixa espectral (Modelo RF 2), aplicando o mesmo pré-processamento em todas as variáveis dessa região espectral (20.433 variáveis). Nesse caso, com aproximadamente 60% dos dados originais foi obtido um modelo com melhor desempenho e em menor tempo (Modelo FD-RF 2).

Tabela 4.1 Resultados dos modelos de classificação para o NAT das 126 amostras de petróleo usando Floresta Randômica e seleção de variáveis.

Modelos	**Pré-processamento**	**Exatidão (%)**					
		RMN de ^{1}H			**RMN de ^{13}C**		
		RF	PCA-RF	FD-RF	RF	PCA-RF	FD-RF
1	Sem	66,7	60,5	73,9	70,8	70,0	73,9
2	SNV	81,6	72,3	87,7	67,5	83,9	73,5
3	Centrado	74,5	65,1	77,5	61,2	62,8	67,4
4	Autoescalamento	74,5	51,3	73,9	54,5	63,0	59,6
5	Derivada	55,6	74,5	71,1	50,0	67,4	53,9

Para o RMN de ^{13}C, o melhor desempenho foi obtido para o modelo construído a partir da redução das variáveis originais pela PCA em associação ao RF (83,9%) e pré-tratamento por SNV (Modelo PCA-RF 2). Em sequência, a combinação de FD e RF apresentou a segunda melhor de exatidão (73,9%), utilizando as 9.900 variáveis selecionadas (Modelo FD-RF 1).

Os resultados baseados nas técnicas de RMN de ^{1}H e ^{13}C mostraram que adotar os métodos de seleção de variáveis combinados a RF foi vantajoso e eficiente. Esse procedimento possibilitou a realização do processo de classificação em menor tempo, que foi verificada pela comparação dos esforços temporais empregados pelo equipamento durante a construção de cada modelo (Tabela D.1, Anexo D). Além disso, o procedimento possibilitou evidenciar a capacidade desses métodos em sumarizar a quantidade de informações contidas nas variáveis originais em um conjunto reduzido de dados que realmente foram relevantes para a discriminação das amostras de petróleo quanto a acidez. Filgueiras et al. (2016)[7] apresentaram semelhante resposta referente à vantagem da utilização do método de identificação e

seleção de variáveis em dados de RMN para petróleos, anteriormente a construção dos modelos não lineares. O pré-tratamento dos dados influenciou positivamente a aquisição das melhores respostas de classificação. Informação que confirma os trabalhos de Filgueiras et al. (2016)[7] e Duarte et al. (2017)[20] quando reportaram o ganho nos resultados dos modelos que foram construídos com os espectros de RMN e pré-processados com SNV.

Além das avaliações quanto ao desempenho, pré-processamento de dados e técnicas aplicadas na construção dos modelos, também foram analisadas e identificadas as variáveis que participaram do processo de construção das árvores de classificação. A Figura 4.4 mostra as variáveis preditoras identificadas nas árvores do modelo RF com exatidão superior a 70% e suas frequências relativas nas regiões espectrais ao longo do espectro de RMN de ^{1}H. Os demais gráficos de frequências dos modelos construídos nos dois modos para RMN de ^{1}H e ^{13}C estão disponíveis no material suplementar (Figuras D.2, D.3 e D.4).

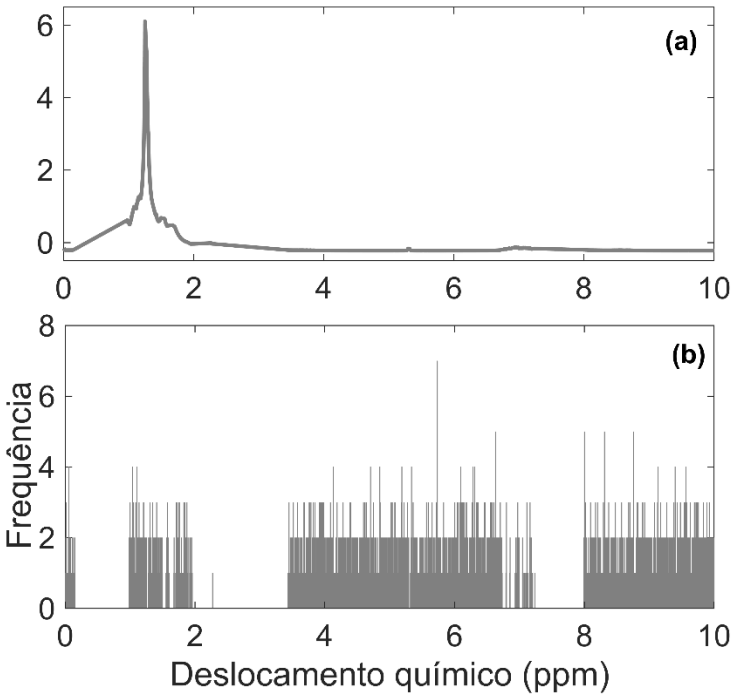

Figura 4.4 Identificação das variáveis preditoras; (a) deslocamento químico no espectro de RMN de ^{1}H (b) gráfico da frequência no deslocamento químico para modelo de classificação FD e RF.

Baseado nessa contribuição, foi possível verificar as regiões espectrais assinalados por Poveda e Molina (2012)[23] que mais influenciaram nos modelos de

classificação. Para os modelos obtidos pela associação da PCA-RF não foi possível aplicar essa análise, em decorrência ao uso das combinações lineares dos dados originais como entrada obtidas através dos componentes principais, impossibilitando a aquisição das informações sobre as variáveis originais.

A identificação de variáveis utilizadas em RF foi realizada nos modelos com a maior exatidão para os sinais do espectro total e para as faixas selecionadas por FD. Os resultados de cada técnica de RMN foram comparados para demonstrar os pesos das variáveis nos dois modos (Modelo RF e Modelo FD-RF). Para o RMN de ^{1}H, mostrou que o Modelo 2, relativo ao pré-processamento por SNV, obteve a melhor exatidão em ambos os modos de análise (Modelo RF 2 e Modelo FD-RF 2, Tabela III). O Modelo RF 2, mostra certa contribuição dos H de aldeídos e dos ácidos carboxílicos (9,0-10,0 ppm), e H aromáticos que estão ligados aos C aromáticos mono, di e poliaromáticos (7,2-9,0 ppm) (Figura 4.5). Em menor proporção estão os H dos grupos aromáticos ligados aos anéis monoaromáticos (6,0-7,2 ppm) e aos H olefínicos (4,5-6,0 ppm). Entretanto, a região com maior aporte de informação para a classificação corresponde aos deslocamentos químicos de 2,0-4,5 ppm, que são relacionados aos H parafínicos e naftênicos, assim como aos CH, CH_2 e CH_3 que estão ligados aos sistemas aromáticos na posição α.

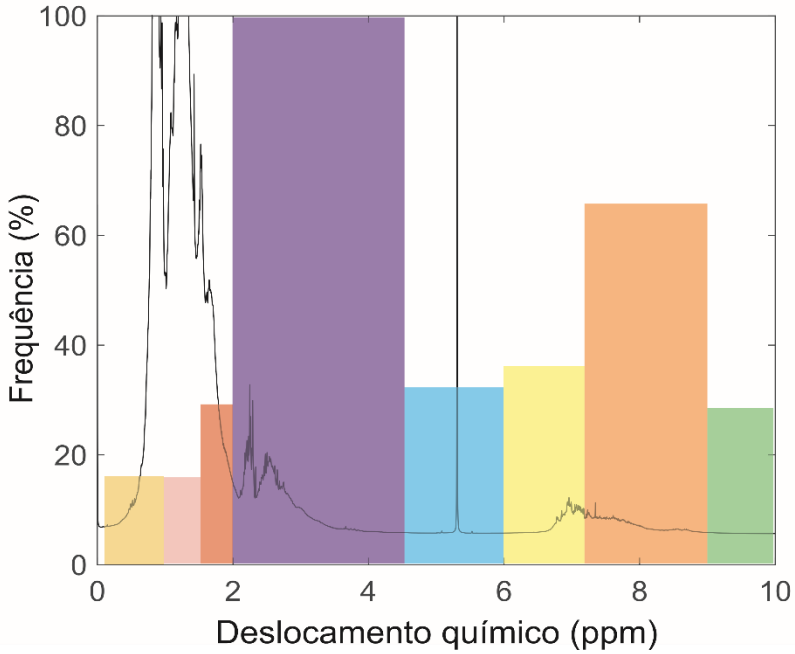

Figura 4.5 Espectro de RMN de ^{1}H e peso das variáveis (frequência relativa) por região espectral para modelo de classificação obtido por RF.

O perfil observado na Figura 4.5, referente à relação entre o peso das variáveis e as regiões, foi modificado após a aplicação do método de seleção (Figura 4.6), devido ao processo para a redução do conjunto original de dados em informações realmente relevantes ter ocasionado uma nova distribuição ao longo do espectro. Nesse sentido, foi possível evidenciar uma maior contribuição das variáveis referentes aos hidrogênios mais ácidos (9.0-10.0 ppm) e aromáticos (6.0-9.0 ppm). Ainda, foi verificada a presença de variáveis com peso na região dos parafínicos, porém em menor proporção quando comparado à faixa de aromáticos.

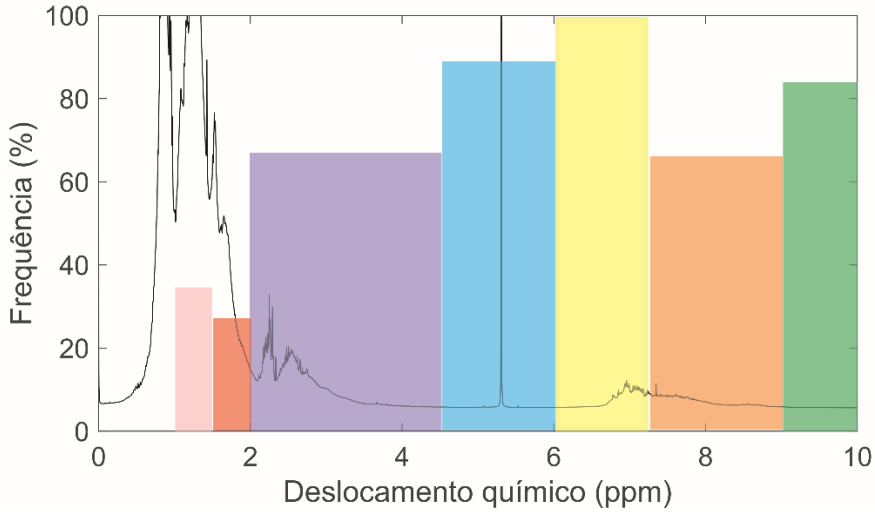

Figura 4.6 Espectro de RMN de ^{1}H e peso das variáveis (frequência relativa) por região espectral para modelo de classificação obtido por FD e RF.

A análise do peso das variáveis no RMN de ^{13}C (Figuras 4.7 e 4.8) apresentou um perfil muito semelhante ao obtido para o Modelo FD-RF 2 com os dados de RMN de ^{1}H (Figura 4.6). O Modelo RF 1 (Figura 4.7), construído com todo o espectro de carbono, apresentou maior expressividade na última região (160-220 ppm), indicando a contribuição dos carbonos de natureza mais ácida na descrição do NAT. Em relação aos carbonos parafínicos (3-78 ppm) e aromáticos (85-160 ppm) observou-se resultados semelhantes para a análise do peso das variáveis. Entretanto, quando foram consideradas as contribuições significativas em função a cada região, verificou-se uma maior representatividade para a faixa dos aromáticos. A região designada aos parafínicos apresentou algumas lacunas com ausência de variáveis importantes para

a acidez e, diferentemente, para os aromáticos foi constatado que toda a faixa espectral possuía contribuições relevantes para a discriminação das amostras.

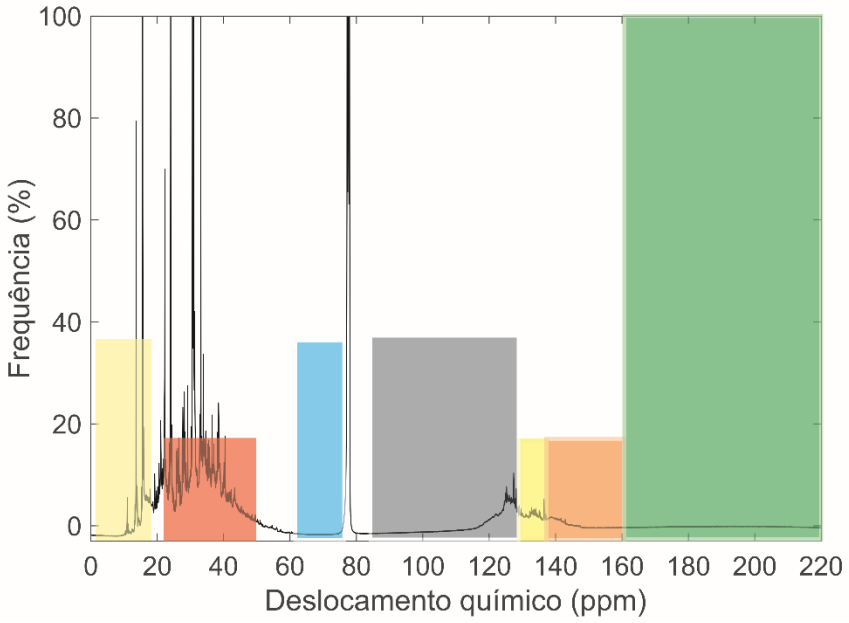

Figura 4.7 Espectro de RMN de ^{13}C e peso das variáveis (frequência relativa) por região espectral para modelo de classificação obtido por RF.

Essas observações são realçadas pelo Modelo FD-RF 1 (Figura 4.8), em que a relação entre as variáveis de maior contribuição dos carbonos ácidos e dos aromáticos foi mais intensa. Nesse sentido, a classificação utilizando o método de seleção por FD, evidenciou a relação das variáveis entre as regiões espectrais do RMN de ^{13}C. Neste trabalho verificamos que o NAT possui uma relação direta e intensa com as regiões designadas aos sinais dos hidrogênios e carbonos de natureza ácida, nos espectros de RMN de ^{1}H e ^{13}C. Em geral, a propriedade apresenta maior relação com a região de aromáticos que de parafínicos. Esse fato ficou mais explícito nos modelos construídos utilizando os métodos de seleção de variáveis.

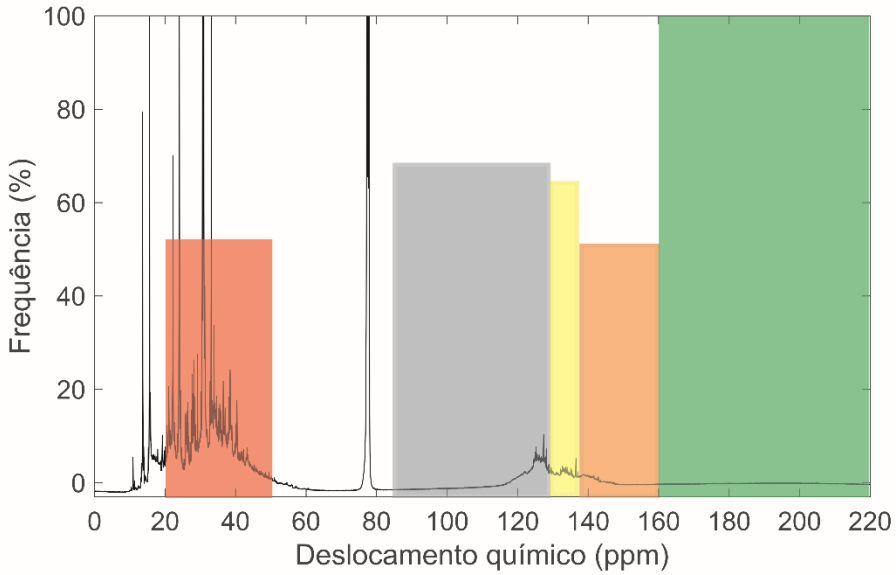

Figura 4.8 Espectro de RMN de ^{13}C e peso das variáveis (frequência relativa) por região espectral para modelo de classificação obtido por FD e RF.

4.5 Conclusões

O método de floresta randômica (RF) é eficaz para discriminar amostras de petróleo em relação ao número de acidez total (NAT). Quando associado a métodos de seleção de variáveis (PCA e FD) adquire maior eficiência, alcançando resultados com exatidão de 88% para dados de RMN de ^{1}H e 84% para dados de RMN de ^{13}C, em menor tempo. Os melhores resultados para as duas técnicas de RMN foram obtidos utilizando SNV como pré-processamento.

Na análise das variáveis de maior importância para discriminar os dados, todos os modelos selecionados mostram uma relação direta da propriedade de interesse com a região ácida nos espectros. De forma geral, a região dos aromáticos apresenta uma contribuição sutilmente maior que a região de parafínicos. Essa última observação é melhor compreendida pelos modelos adquiridos com seleção de variáveis por FD. Nesse sentido, o método de discriminante de Fisher é eficiente em identificar as variáveis com maior relação com a propriedade físico-química de NAT e mostra ser capaz de enfatizar os resultados adquiridos com toda a faixa espectral. O NAT possui uma relação direta com a região ácida dos sinais de hidrogênio e carbono, além disso, a propriedade possui uma maior relação com a região de aromáticos que parafínicos. Esse fato é potencializado nos modelos construídos com métodos de seleção de variáveis.

4.6 Referências Bibliográficas

1. LIU, W.; LIU, C.; HU, X.; YANG, J.; ZHENG, L. Application of terahertz spectroscopy imaging for discrimination of transgenic rice seeds with chemometrics, *Food Chemistry*. **210**, 415–421, 2016.

2. LEE, S.; CHOI, H.; CHA, K.; CHUNG, H. Random forest as a potential multivariate method for near-infrared (NIR) spectroscopic analysis of complex mixture samples: Gasoline and naphtha. *Microchemical Journal*, **110**, 739–748, 2013.

3. CANIZO, B. V.; ESCUDERO, L. B.; PÉREZ, M. B.; PELLERANO, R. G. Intra-regional classification of grape seeds produced in Mendoza province (Argentina) by multi-elemental analysis and chemometrics tools. *Food Chemistry*, **242**, 272–278, 2018.

4. LOVATTI, B. P. O.; NASCIMENTO, M. H. C.; NETO, A. C.; CASTRO, E. V. R.; FILGUEIRAS, P. R. Use of random forest in the identification of important variables. *Microchem. J*., **145**, 129-1134, 2019.

5. MASILI, A., PULIGHEDDU, S. SASSU, L. SCANO, P.; LAI, A. Prediction of physical-chemical properties of crude oils by ^{1}H NMR analysis of neat samples and chemometrics. *J. Chemometr*., **50,** 729-738, 2012.

6. FILGUEIRAS P. R.; TERRA, L.; CASTRO, E. V. R.; OLIVEIRA, L. M. S. L.; DIAS, J. C. M.; POPPI, R. J. Prediction of the distillation temperatures of crude oils using ^{1}H NMR and support vector regression with estimated confidence intervals. *Talanta*, **142**: 197-205, 2015.

7. FILGUEIRAS, P. R.; PORTELA, N. A.; SILVA, S. R. C.; CASTRO, E. V. R.; OLIVEIRA, L. M. S. L.; DIAS, J. C. M.; NETO, A. C.; ROMÃO, W.; POPPI, R. J. Determination of saturates, aromatics, and polars in crude oil by ^{13}C NMR and support vector regression with variable selection by genetic algorithm. *Energy Fuels*. **30,** 972-1978, 2016.

8. BREIMAN, L. Random Forests. *Machine Learning*, **45**, 5-32, 2001.

9. SHEN, T.; KONG, W.; LIU, F.; CHEN, Z.; YAO, J.; WANG, W.; PENG, J.; CHEN, H.; HE, Y. Rapid Determination of Cadmium Contamination in Lettuce Using Laser-Induced Breakdown Spectroscopy. *Molecules*, **23**, 2930-2945, 2018.

10. AMJAD, A.; ULLAH, R.; KHAN, S.; BILAL, M.; KHAN, A. Raman spectroscopy based analysis of milk using random forest classification. *Vibrational Spectroscopy*, **99**, 124–129, 2018.

11. SPEIGHT, J. G.; Handbook of Petroleum Analysis, John Wiley and Sons Inc: New Jersey, **2001**, 519p.

12. ASTM D664-11. Standard Test Method for Acid Number of Petroleum Products by Potentiometric Titration. ASTM International. West Conshohocken, PA; 2011.

13. SAD, C. M. S.; LACERDA, J. R. V.; FILGUEIRAS, P. R.; RIGONI, V. S.; BASSANE, J. F. P.; CASTRO, E. V. R.; PEREIRA, K. S.; SANTOS, M. F. P. Limitations of the pour point measurement and the influence of the oil composition on its detection using principal component analysis, *Energy Fuels*, **28**, 1686–1691, 2014.

14. SAVORANI, F.; TOMASI, G.; ENGELSEN, S. B. Icoshift: A versatile tool for the rapid alignment of ^{1}D NM spectra, *J. Magn. Resson.,* **202**, 190-202, 2010.

15.TOMASI, G.; SAVORANI, F.; ENGELSEN, S. Icoshifit: an effective tool for the alignment of chromatographic data. *J. Chromatogr.,* **1218**, 7832-7840, 2011.

16. BLANCO, M.; COELHO, J.; ITURRIAGA, H.; MASPOCH, S.; PEZUELA, C. Effect of data preprocessing methods in near-infrared diffuse reflectance spectroscopy for the determination of the active compound in a pharmaceutical preparation. *Appl. Spectrosc.*, **51**, 240-246, 1997.

17. GELADI, P.; KOWALSKI, B. R. Partial least-squares regression: a tutorial. *Anal. Chim. Acta*, **185**, 1-17, 1986.

18. BARNES, R. J.; DHANOA, M.S.; LISTER, S. J. Standard Normal Variate transformation and De-trending of near-infrared diffuse reflectance spectra. *Appl. Spectrosc*. **43**, 772-777, 1989.

19. FEARN, T.; RICCIOLI, C.; GARRIDO-VARO, A.; GUERRERO-GINEL, J. E. On the geometry of SNV and MSC, *Chemometrics intel. Lab. Syst.* **96**, 22-26, 2009.

20. DUARTE, L. M.; FILGUEIRAS, P. R.; DIAS, J. C. M.; OLIVEIRA, L. M. S. L.; CASTRO, E. V. R.; OLIVEIRA, M. A. L. Study of Distillation Temperature Curves from Brazilian Crude Oil by 1 H Nuclear Magnetic Resonance Spectroscopy in Association with Partial Least Squares Regression. *Energy Fuels*, **31**, 3892-3897, 2017.

21. PIERCE, K. M.; PARSONS, B. A.; SYNOVEC, R. E. Pixel-Level Data Analysis Methods for Comprehensive Two-Dimensional Chromatography. In: Fundamentals and Analytical Applications of Multiway Calibration. *Elsevier*, **1**, 427-463, 2015.

22. JOHNSON, K. J.; SYNOVEC, R. E. Pattern recognition of jet fuels: comprehensive GC×GC with ANOVA-based feature selection and principal component analysis. *Chemom. Intell. Lab. Syst.*, **60**, 225-237, 2012.

23. POVEDA, J. C.; MOLINA, D. R. Average molecular parameters of heavy crude oils and their fractions using NMR spectroscopy. *Journal of Petroleum Science and Engineering*, **84**, 1–7, 2012.

24. GRUBERI, L. D. A; DAMASCENOI, F. C.; CARAMÃO, E. B.; JACQUESI, R. A.; GELLERII, A. M.; CAMPOS, M. C. V. Ácidos naftênicos no petróleo. *Química Nova*. **35**, 1423-1433, 2012.

25. BERTHEUSSEN, A.; SIMON, S.; SJÖBLOM, J. Equilibrium Partitioning of Naphthenic Acid Mixture Part 2: Crude Oil-Extracted Naphthenic Acids. *Energy Fuels.* **32**, 9142-9158, 2018.

5. CONCLUSÃO GERAL

A utilização da Quimiometria associada as técnicas espectroscópicas tem mostrado potencial em resolver problemas qualitativos. Grandes quantidades de variáveis podem adquirir significado químico com o uso de métodos multivariados na classificação desses dados.

O uso de Floresta Randômica (RF) como um método de classificação demonstra ser atrativo por solucionar problemas de natureza complexa, como petróleo e avaliar a importância das variáveis. Com a identificação das variáveis de maior contribuição no modelo é possível obter uma interpretação química baseada nas informações moleculares da técnica de RMN.

A associação de métodos de seleção de variáveis e RF em dados de RMN mostra ser vantajoso por produzir modelos robustos em menor tempo. Esse artifício permite que grandes conjuntos de dados sejam classificados de forma rápida e eficiente.

Os métodos de aprendizagem de máquina PCA e k-NN são uma alternativa metodológica eficiente para a identificação de perfis de petróleo. A similaridade química entre as amostras pode ser obtida de forma rápida por meio de dados de MIR, baseada na impressão digital da molécula, e métodos quimiométricos.

ANEXOS

Anexo A – Exemplificação sobre a construção de uma árvore de decisão pelo método CART

Para esse exemplo é utilizado um conjunto de dados sobre três espécies da flor de íris (150 amostras) com o objetivo de elucidar melhor como a árvore de decisão particiona o espaço em sub-regiões. Nesse caso, para obter uma visualização mais fácil dos dados em um espaço bidimensional são utilizados dois atributos: largura e comprimento da pétala da flor (Figura A.1).

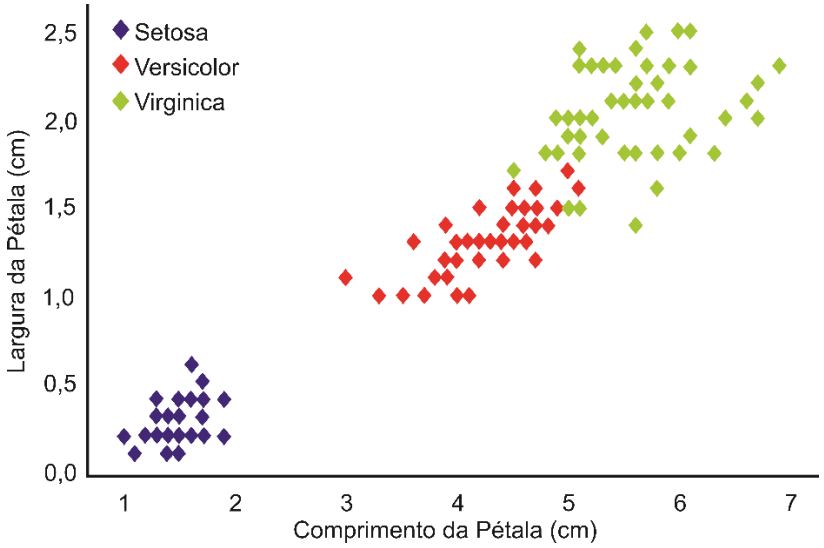

Figura A.1 Gráfico bidimensional (largura e comprimento) das espécies da flor de íris de Fisher.

O gráfico bidimensional (Figura A.1) apresenta as três espécies de flor de íris representadas por diferentes cores: setosa (azul), versicolor (vermelha) e virginica (verde). Tendo isso em vista, o objetivo é construir uma árvore de decisão para determinar a qual espécie uma nova flor de íris pertence dada suas características (comprimento e largura da pétala). Para isso, o espaço é particionado recursivamente, por meio de cortes ortogonais, levando em consideração uma variável por vez, que maximiza a pureza das sub-regiões resultantes, ou seja, a variável que melhor representa a homogeneidade de cada espécie ou grupo). A cada seleção da melhor variável preditora proporciona a divisão do espaço e esse processo é representado por um nó na árvore de decisão. A primeira divisão (nós raiz da árvore), leva em consideração todos os exemplos (pontos) do espaço ao encontrar o ponto de corte (melhor variável preditora) que maximiza a pureza, de acordo com critério de impureza

(índice de Gini), das sub-regiões resultantes. Supondo que o comprimento da pétala de valor 2,45 cm (Figura A.2) é o ponto de corte que melhor divide o espaço. Portanto, essa variável pertence ao nó raiz da árvore de decisão.

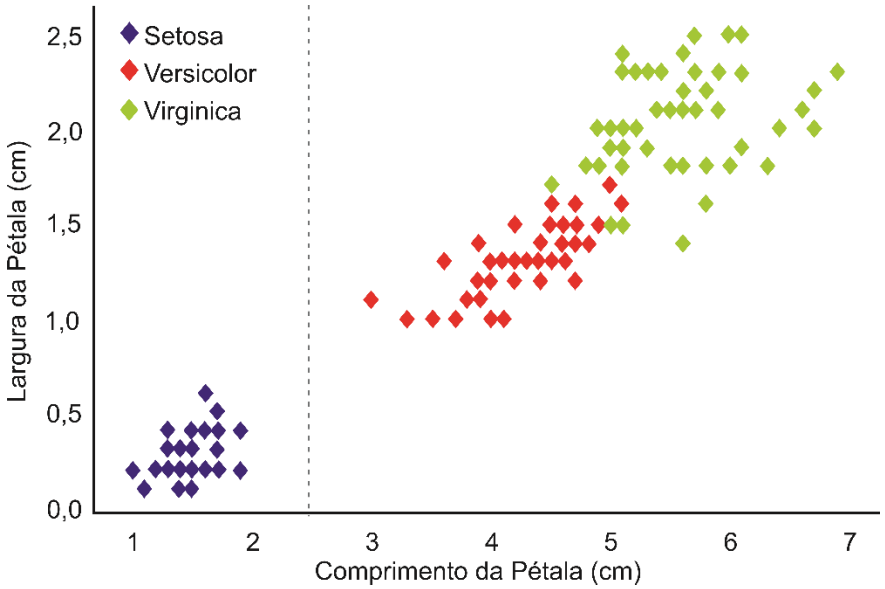

Figura A.2 Gráfico bidimensional contendo a primeira partição do espaço.

O gráfico (Figura A.2) mostra como fica após o particionamento para o nó raiz. De fato, a partir desse corte é obtida uma região totalmente pura (separa a classe setosa das demais) e uma outra heterogênea. Durante a construção de uma árvore de decisão, ao alcançar uma região pura, surge um nó folha contendo o rótulo da classe, nesse caso classe setosa (Figura A.3). No nó folha são cessados os particionamentos, já a região a direita do ponto de corte não é pura, assim, os particionamentos continuam até que se alcance sub-regiões puras contendo a classificação das demais espécies (100 amostras) (Figura A.4).

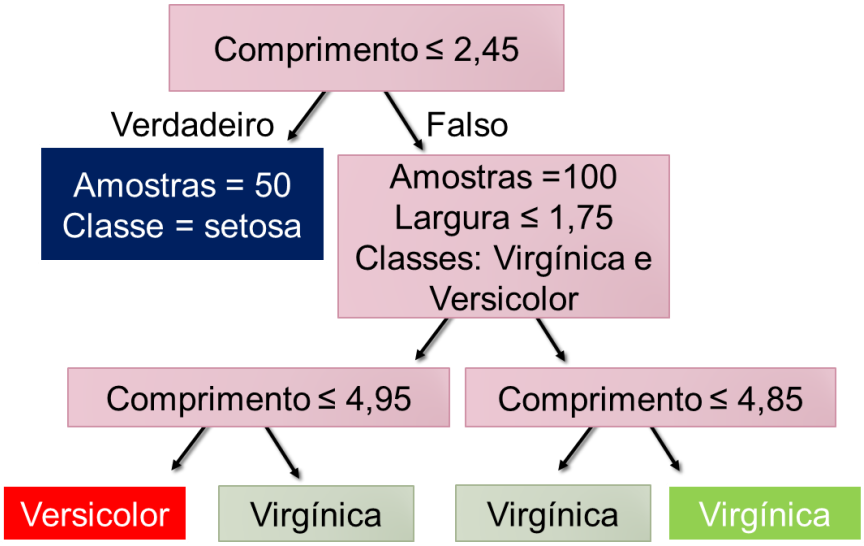

Figura A.3 Construção da árvore de decisão.

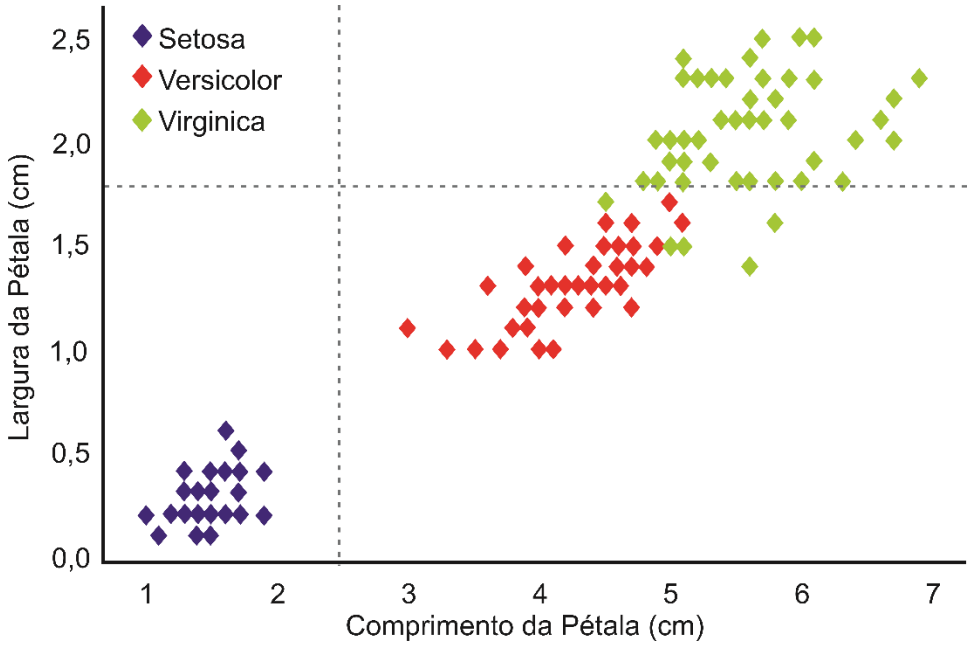

Figura A.4 Gráfico bidimensional contendo a segunda partição do espaço.

Nesse sentido, outra partição foi feita no eixo da largura da pétala no valor 1,75 cm (Figura A.4). Como pode-se notar, foram geradas mais duas novas sub-regiões disjuntas, que, por sua vez, não estão completamente puras. Esse processo é repetido recursivamente até que um critério de parada seja alcançado ou que todas as folhas sejam puras, que implica na construção até sua profundidade máxima (Figura A.3).

Anexo B – Material suplementar da aplicação 1: uso de Floresta Randômica na identificação de variáveis importantes

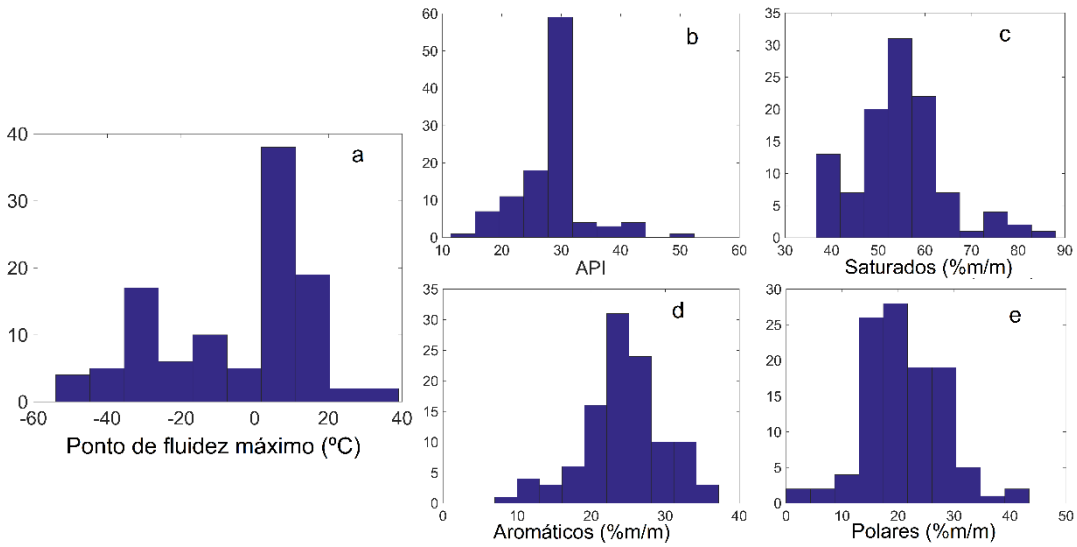

Figura B.1 Histograma de propriedades físico-químicas (gravidade API, teor de saturados, teor de aromáticos e teor de polares) para as 105 amostras de petróleo.

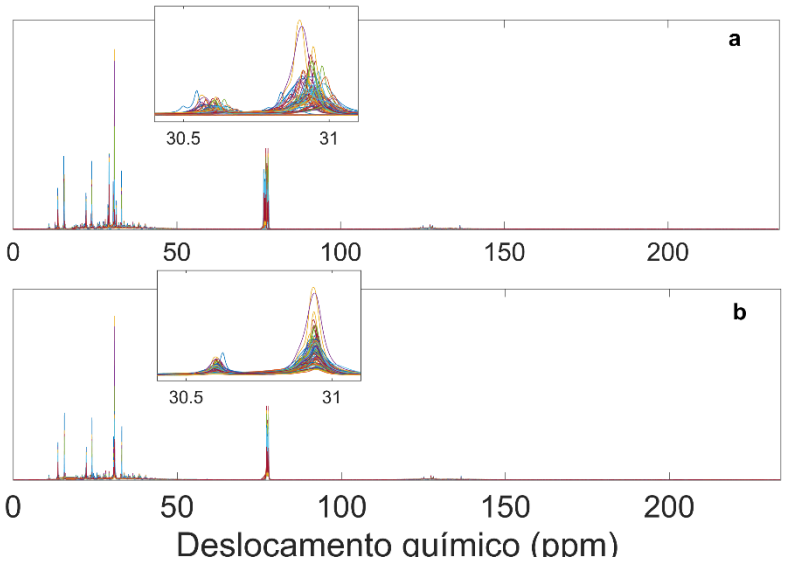

Figura B.2 Espectro de RMN de ^{13}C das 105 amostras de petróleo usadas nos modelos de classificação. (a) dados originais (b) após alinhamento pelo algoritmo icoshift.

Tabela B.1 Coeficiente de correlação entre propriedades: API, teor de saturados (SAT), teor de aromáticos (ARO) e ponto de fluidez máximo (PFM).

	API	**SAT**	**ARO**	**PFM**
API	1,0000	0,7853	-0,6260	0,2025
SAT	0,7853	1,0000	-0,6646	0,0839
ARO	-0,6260	-0,6646	1,0000	-0,4013
PFM	0,2025	0,0839	-0,4013	1,0000

Tabela B.2 P-valor no teste de hipótese para significância estatística do cálculo dos coeficientes de correlação. H^0 = não há evidência significativa da correlação entre propriedades.

	API	**SAT**	**ARO**	**PFM**
API	1,0000	0,0000	0,0000	0,0356
SAT	0,0000	1,0000	0,0000	0,3882
ARO	0,0000	0,0000	1,0000	0,0000
PFM	0,0356	0,3882	0,0000	1,0000

Anexo C – Tabelas de assinalamentos estruturais de Molina e Poveda

Tabela C.1 Assinalamentos estruturais de RMN de ^{1}H.

Região (ppm)	Designação estrutural
0,1-1,0	H parafínico do tipo CH_3, em posição γ ou mais em relação ao sistema aromático
1,0-1,5	H parafínico associados a sistemas aromáticos na posição β e alquílicos terminais
1,5-2,0	H naftênico do tipo CH_2 ligados a sistemas aromáticos na posição β
2,0-4,5	H parafínico e naftênico do tipo CH, CH_2 e CH_3 ligados a sistemas aromáticos na posição α; outros grupos -OH e -SH
4,5-6,0	H olefínicos
6,0-7,2	H aromático ligado a anéis monoaromáticos
7,2-9,0	H aromático ligado a C aromáticos de anéis di e poliaromáticos
9,0-12,0	H de aldeído e ácido carboxílico

Tabela C.2 Assinalamentos estruturais de RMN de ^{13}C.

Região (ppm)	Designação estrutural
3,0-18,5	C do tipo CH_3
18,5-21,5	C metílico em posição α em relação ao anel aromático
21,5-50,0	C parafínicos e naftênicos do tipo CH e CH_2
50,0-60,0	C parafínico terciários (CH) e quatenários (C)
60,0-78,0	C parafínico e naftênico em posição α ao OH
85,0-129,2	C aromático olefínico protonado, C em sistemas benzonaftênicos e alquílicos e C em junção de três anéis aromáticos
129,2-137,0	C aromático ligado a grupos metil e C cabeça de ponte entre anéis aromáticos
137,0-160,0	C aromático não protonado
160,0-220,0	C de cetonas, ácidos, aldeídos e quinonas

Anexo D – Material suplementar da aplicação 2: Diferentes estratégias para o uso de floresta Randômica em dados de Ressonância Magnética Nuclear

Tabela D.1 Tempo gasto na execução dos modelos de classificação

Modelo	Pré-processamento	Tempo (Horas)					
		RMN de ^{1}H			RMN de ^{13}C		
		RF	PCA-RF	FD-RF	RF	PCA-RF	FD-RF
1	Sem	0,48	0,29	0,31	2,32	0,96	0,49
2	SNV	0,47	0,29	0,35	2,29	0,95	0,62
3	Centrado	0,50	0,29	0,28	2,31	0,94	0,30
4	Autoescalamento	0,51	0,29	0,27	2,33	0,65	0,55
5	Derivada	0,45	0,29	0,21	2,30	0,94	0,43

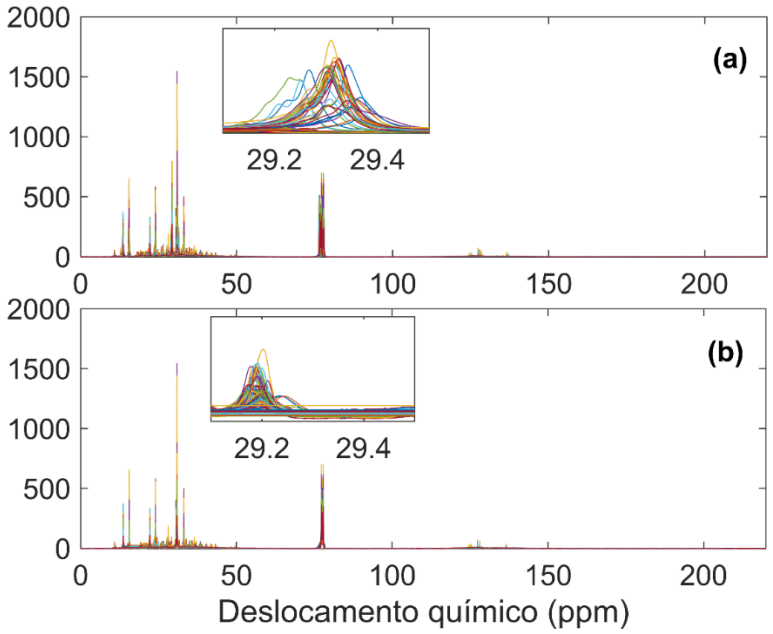

Figura D.1 Espectros de RMN de ^{13}C das 126 amostras de petróleos utilizadas nos modelos de classificação. (a) dados originais e (b) após o alinhamento pelo algoritmo icoshift.

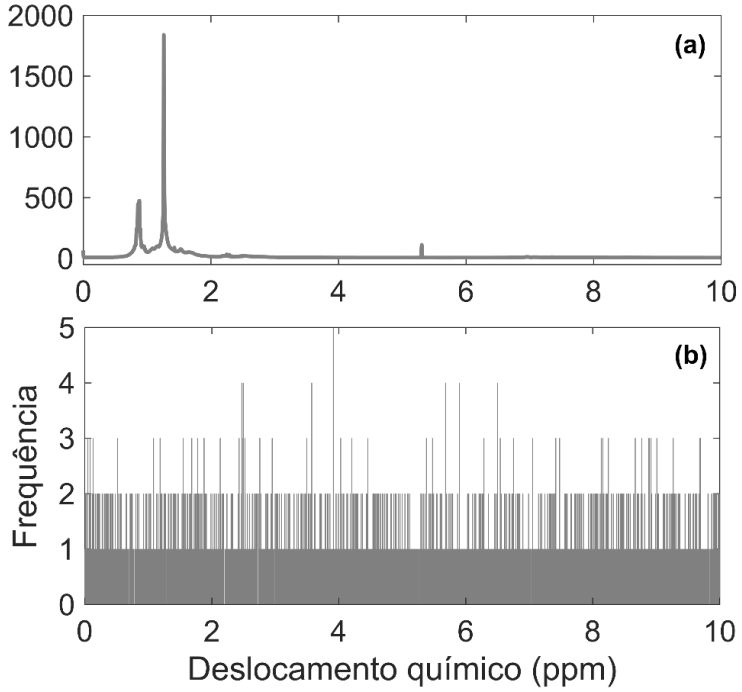

Figura D.2 Identificação das variáveis preditoras; (a) deslocamento químico no espectro de RMN de ^{1}H (b) gráfico da frequência no deslocamento químico para modelo de classificação RF.

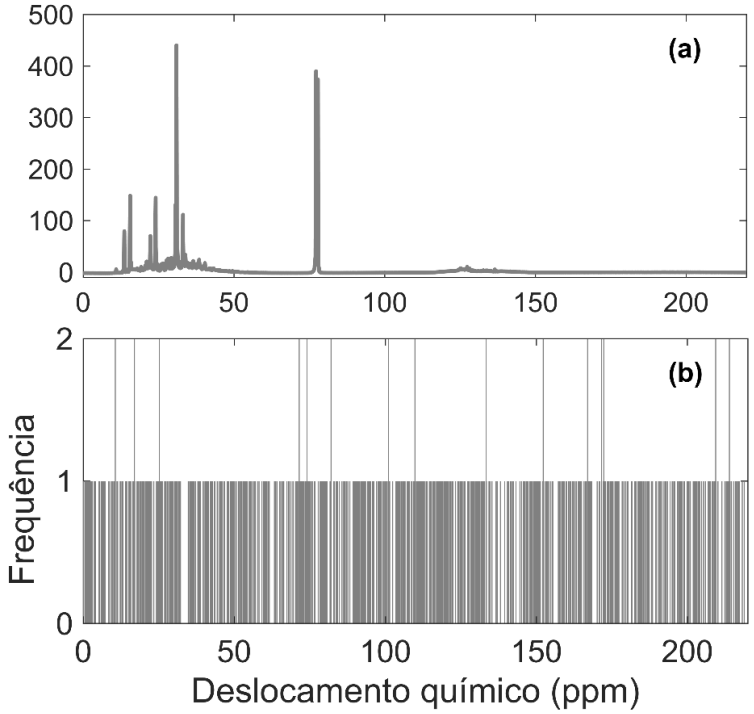

Figura D.3 Identificação das variáveis preditoras; (a) deslocamento químico no espectro de RMN de ^{13}C (b) gráfico da frequência no deslocamento químico para modelo de classificação RF.

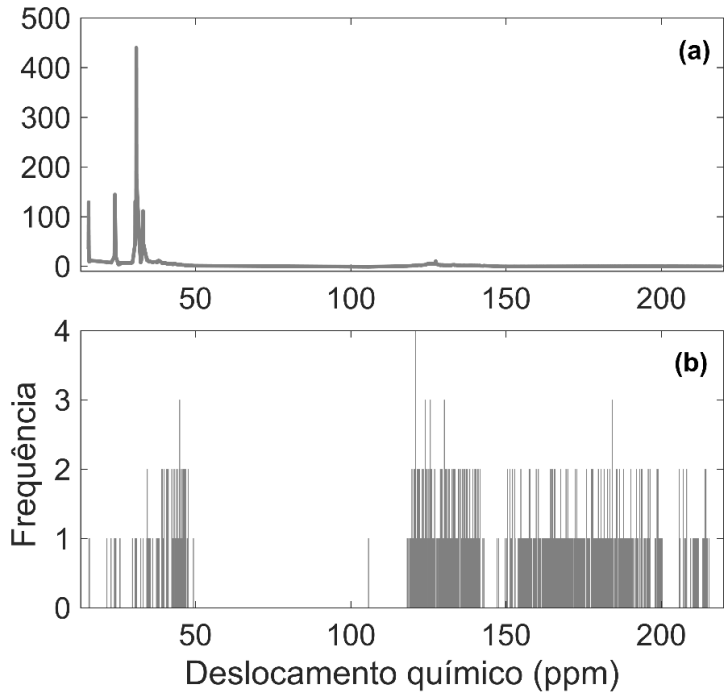

Figura D.4 Identificação das variáveis preditoras; (a) deslocamento químico no espectro de RMN de ^{13}C (b) gráfico da frequência no deslocamento químico para modelo de classificação FD e RF.

Anexo E – Material suplementar da aplicação 3: Identificação de perfis de petróleo por espectroscopia no infravermelho e quimiometria.

Tabela E.1 Propriedades físico-químicas das amostras de treinamento e teste.

Amostra/Parâmetro	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13
A1	41,9	62,8	6	23,35	3,15	1,95	71,9	16,2	11,8	0,1	0,171	716	0,17
A2	17	11,3	-9	14,52	485,9	7,32	41,6	29,8	27,17	1,43	0,562	3892	3,09
A3	25,5	28,9	-54	14,29	21,55	3,26	58,6	23,4	18	0	0,351	2717	1,18
A4	29,9	29,6	-27	12,78	9,658	3,1	52,7	33,6	12,6	0,97	0,216	2800	0,07
A5	26,8	32,7	9	20,76	31,35	4,35	53,8	24,1	22,1	0	0,412	3529	0,29
A6	26,7	43,4	-12	19,04	20,88	4,75	56	26,9	16	1,1	0,5	2960	0,19
A7	28,4	40,8	9	20,13	17,84	4,3	51,8	21,9	25,6	0,75	0,308	3355	0,17
A8	29,4	48,2	6	17,03	14,61	3,5	58,3	21,9	19,8	0	0,36	2535	0,35
A9	29,6	48,2	12	19,73	14,1	3,46	60,3	22,7	17	0	0,361	2527	0,24
A10	28,7	41,1	6	19,59	17,93	4,11	58,4	19,4	22,2	0	0,385	2836	0,51
A11	29,2	45,6	12	19,31	12,6	3,17	54	24	22	0	0,352	2800	0,27
A12	32,1	51	3	24	8,9784	3,11	56,6	24,4	19	0	0,266	1300	0,45
A13	37,8	50,9	-18	20,11	4,126	2,75	74,1	13,9	11,2	0,75	0,169	1094	0,73
A14	30,2	50,4	6	17,9	12,2	2,8	63,1	18,2	18,6	0,1	0,353	2401	0,32
A15	41	29	-6	19,52	2,7646	0,82	80,8	12,8	6,4	0	0,0838	579	0,03
A16	25,9	50,6	-15	17,54	23,77	5,7	48,6	22,2	26,6	2,6	0,19	3674	0,815
A17	35,9	58,8	12	18,87	5,552	1,2	46	35,1	18,9	0	0,04	1312	0,5

A18	23,6	42,5	21	25,32	110	6,5	40,4	16,2	42,75	0,65	0,556	4009	0,42
A19	36,7	29,4	15	25,48	6,1578	2,92	76,6	13,5	9,3	0,6	0,175	1060	0,13
A20	22,7	23,2	-51	13,96	44,3278	5,89	40,2	37,2	20,8	1,94	0,68	3900	1,16
A21	27,8	48,5	12	19,73	20,5263	4,09	55,1	27,3	17,05	0,55	0,36	3500	0,3
A22	30,6	34,9	3	20,19	12,6595	2,95	61,5	23,1	15,35	0,05	0,278	2400	0,11
A23	29,6	45,6	12	19,31	12,6	3,17	54	24	22	0	0,352	2800	0,27
A24	25,9	50,6	-15	17,54	23,77	5,7	48,6	22,2	26,6	2,6	0,815	3674	0,19
A25	25,6	28,9	-54	14,29	21,55	3,26	58,6	23,4	18	0	0,351	2717	1,18
A26	29,4	48,2	6	17,03	14,61	3,5	58,3	21,9	19,8	0	0,36	2535	0,35
A27	28,7	34,4	-42	14,12	13,06	4,25	57	25,5	15,3	2,2	0,3	3166	0,16
A28	31,7	73	15	36,53	9,5587	3,9	57,1	26,2	14,7	2,01	0,155	2300	0,03
A29	26,8	32,7	9	20,76	31,35	4,35	53,8	24,1	22,1	0	0,412	3529	0,29
A30	23,6	42,5	21	25,32	110	6,5	40,4	16,2	42,8	0,65	0,556	4009	0,42
A31	28,4	40,8	9	20,13	17,84	4,3	51,8	21,9	25,6	0,75	0,308	3355	0,17
A32	41	29	-6	19,52	2,7446	0,82	80,8	12,8	6,4	0	0,0838	579	0,03
A33	29	50,6	6	19,3	15,29	3,29	52,1	30,1	17,8	0	0,336	2912	0,37
A34	30,8	41,3	6	20,69	10,87	2,64	62,6	18,7	18,8	0	0,286	2618	0,09
A35	29,6	59,3	-15	20,23	17,19	4,27	55,6	19,2	24,6	0,5	0,381	3215	0,27
A36	28,3	45,5	9	20,9	17,25	4,26	58	25,9	15,4	0,7	0,305	2951	0,15
A37	28,1	48,7	-24	20,46	20,28	4,35	59,7	23,4	16,1	0,7	0,37	3152	0,24
A38	22,1	28	-33	16,22	53,57	5,27	41,3	31,3	26,45	0,95	0,663	3316	1,32
A39	28,1	36,4	-33	21,19	19,6	3,8	59,4	21	19,6	0	0,35	3200	0,2

A40	28,8	29,8	12	20,79	16,88	3,49	55,3	27,4	0	0	0,368	3077	0,23
A41	29,7	35,3	3	19,78	11,89	3,77	57,7	26,1	15,4	0,8	0,307	2401	0,12
A42	22,2	32,2	-39	12,18	37,05	5,33	49,8	29,2	19,6	1,4	0,686	3500	1,25
A43	31,4	24,9	-12	35,49	9,707	2,44	66,2	20,2	13,6	0	0,336	2192	0,35
A44	30,5	48,7	-24	19,17	12,83	2,93	57,7	27,3	15	0	0,351	2366	0,26
A45	30,9	47,4	-3	20,14	10,97	2,3	59	24,4	16,5	0,1	0,295	2100	0,09
A46	30,2	37,8	9	17,51	11,39	3,58	58,5	23,6	17,4	0,5	0,315	2362	0,27
A47	17,7	22,4	-24	14,92	305,1	7	36,7	27,6	33,4	2,2	0,609	4200	3,07
A48	32,4	20,2	-18	19	8,374	2,01	64,8	22,7	12,4	0,1	0,301	2439	0,22
A49	29,1	49,5	-27	14,73	11,68	3,72	53,9	29,2	16	0,9	0,303	2765	0,16
A50	37,8	50,9	-18	20,11	4,126	2,75	74,1	13,9	11,2	0,75	0,169	1094	0,73
A51	26,8	32,7	9	20,76	31,35	4,35	53,8	24,1	22,1	0	0,412	3529	0,29
A52	27,9	20,7	-3	14,64	12,32	4,57	56,3	25,2	16,9	1,6	0,394	3147	0,21
A53	28	30,1	3	19,39	19,01	3	59,8	24,2	16	0	0,328	2813	0,31
A54	36,7	29,4	15	25,48	6,077	2,92	76,6	13,5	9,3	0,6	0,175	1060	0,13
A55	41	29	-6	19,52	1,7446	0,82	80,8	12,8	6,4	0	0,0838	579	0,03
A56	29	50,6	6	19,3	15,29	3,29	52,1	30,1	17,8	0	0,336	2912	0,37
A57	26,7	43,4	-12	19,04	20,88	4,75	56	26,9	16	1,1	0,5	2960	0,19
A58	28,4	40,8	9	20,13	17,84	4,3	51,8	21,9	25,6	0,75	0,308	3355	0,17
A59	30,6	25,6	-33	17,3	8,989	3	59,7	26,2	13,5	0,6	0,177	1820	0,04
A60	28,8	57,25	9	20,74	17,14	4,33	50,8	22,9	25,6	0,7	0,368	3279	0,27
A61	29,8	50,2	12	20,83	15,16	3,88	54,7	21,4	24	0	0,327	2630	0,13

A62	28,9	56,8	3	21,17	19,46	4,35	52	20,5	26,7	0,8	0,373	3208	0,29
A63	18,3	18,9	-30	15,13	156,3	7,28	39,4	31,5	26,4	2,6	0,797	4302	1,43
A64	31,3	58,5	-9	20,48	11,07	2,63	57,5	23,7	18,8	0	0,29	2399	0,13
A65	26,9	40,3	-9	17,67	19,63	4,72	48,2	26,2	24	1,5	0,753	3288	0,12
A66	30,6	40,9	-9	22,37	10,07	2,67	63,9	21,6	14,4	0,1	0,3	2460	0,11
A67	29,3	48,5	-27	21,48	12,63	4,67	56,6	22,7	19,8	0,8	0,74	1600	0,29
A68	28,4	52,6	-18	18,69	16,7	4,21	51	21,4	26,7	0,9	0,324	3141	0,28
A69	30,3	36	-12	20,19	10,86	3,74	64,1	17,8	17,2	0,9	0,306	2274	0,06
t1	40,5	38,5	18	23,25	4,186	2,23	74,7	14,8	10,6	0	0,181	757	0,14
t2	21,9	26,1	-33	12,14	40,24	5,28	45,2	33,5	19,5	1,2	0,743	3674	1,31
t3	38,5	52,4	9	17,86	4,205	1,2	82	17,8	0,2	0	0,0863	924	0,08
t4	46,4	45,4	-21	19,46	1,9818	0,29	80,8	8,5	10,7	0	0,0518	76	0,04
t5	17,7	10,06	-21	15,54	221,2	6,35	41,3	34,5	22,6	1,6	0,633	4200	2,78
t6	28,5	62,8	9	21,8	17,86	4,36	53,4	24,7	21,3	0,6	0,383	3162	0,3
t7	22	16,6	-30	14,48	60,24	6,03	44,2	26,5	27,2	2,1	0,555	3747	1,73
t8	22,6	34,3	-36	12,73	33,04	3,97	48,1	33,3	17,75	0,7	0,701	3500	1,16
t9	17,6	11,8	-15	16	266,3	8,15	38,3	31,4	26	4,2	0,727	4817	1,85
t10	24,4	26,3	-30	17,58	37,03	5,7	44,8	22,8	31	1,4	0,51	3477	0,58
t11	30,8	36,4	9	19,88	10,48	3,71	59,8	19,4	20,3	0,5	0,306	2220	0,05
t12	26,8	32,7	9	20,76	31,35	4,35	53,8	24,1	22,1	0	0,412	3529	0,29

* Os parâmentros P1 a P13 representam as propriedades físico-químicas listadas na Tabela 5.1.

Anexo F: Aplicação 3: Artigo publicado na *Fuel*

LOVATTI, B. P. O.; SILVA, S. R. C.; PORTELA, N. A.; SAD, C. M. S.; RAINHA, K. P.; ROCHA, J. T. C.; ROMÃO, W.; CASTRO, E. V. R.; FILGUEIRAS, P. R. Identification of petroleum profiles by infrared by spectroscopy and chemometrics. Fuel, **264**, 1-7, 2019.

APLICAÇÃO 3: IDENTIFICAÇÃO DE PERFIS DE PETRÓLEO POR ESPECTROSCOPIA NO INFRAVERMELHO E QUIMIOMETRIA

1. Introdução

O petróleo é uma mistura de ocorrência natural consistindo predominantemente de hidrocarbonetos e apresentando quantidades menores de compostos sulfurados, nitrogenados e oxigenados. Os diversos compostos que o constituem, combinados de forma variável, conferem-lhe características diferenciadas.[1-3] Estudos mais aprofundados sobre a composição de petróleo foram motivados devido às necessidades de caracterização geoquímica dos reservatórios nos quais geólogos e geoquímicos estudam a relação do petróleo a sua rocha geradora e ao ambiente composicional, além de processos de migração e estudos genéticos, do processamento primário e refino.[1,4,5] A composição química do petróleo é dependente da matriz orgânica que lhe deu origem e do processo de formação[6], assim, óleos oriundos de diferentes poços podem apresentar alterações significativas na composição.[7] A indústria petrolífera tem interesse em conhecer as propriedades físico-químicas do óleo antes de sua exploração a fim de estimar custos de processamento, transporte, armazenamento e refino.[5,8-11]

Para classificar petróleo de acordo com sua origem é essencial o conhecimento prévio de algumas propriedades físico-químicas.[12-14] Essas propriedades são determinadas a partir de uma série de ensaios nos quais alguns podem ser laboriosos, demandam muito tempo ou grandes quantidades de amostra.[4,10] Neste processo, quanto maior o número de ensaios realizados sobre a amostra, melhor a qualidade da estimativa sobre o óleo em estudo. Entretanto, essa estratégia pode exigir um longo tempo para determinar todos os parâmetros físico-químicos necessários para a caracterização do petróleo, volume de amostra e recursos financeiros, retardando tomadas de decisões também sobre a produção e refino.[1,6,10,11,15,16]

Estudos alternativos de caracterização e avaliação das propriedades físicas e químicas do petróleo têm mostrado potencialidade na associação de técnicas espectroscópicas (infravermelho na região do médio e próximo) com quimiometria.[17-23] Técnicas espectroscópicas permitem aquisição de grande quantidade de informação sobre a amostra, a nível molecular. Entretanto, a extração dessa informação química por procedimentos univariados se tornaria laborioso ou impossível. Nesse contexto, faz-se necessário o emprego de técnicas quimiométricas.[23-25] Dentre as principais vertentes da Quimiometria, destaca-se o reconhecimento de padrões[26] que tem por finalidade discriminar e encontrar padrões em dados de origem multivariada buscando seu agrupamento por similaridades. Estes padrões relacionam a identidade de uma amostra com suas características químicas.[27]

Os métodos de reconhecimento de padrões não supervisionados mais aplicados em química são: a Análise de Componentes Principais (PCA) e o k-Vizinhos mais Próximos (k-NN)[23]. Urdal et al. (1986)[28] fizeram uso de k-NN para agrupar amostras de petróleo de acordo com semelhanças de seus parâmetros físico-químicos comparando seus resultados com outros métodos quimiométricos. Chiaberge et al. (2013)[29] classificaram amostras de petróleo bruto extraídas de diferentes poços através da PCA (APPI FTICR MS). Altinpinar et al. (2013)[30], discrimina solos contaminados com diferentes hidrocarbonetos de petróleo com auxílio de micro-NIR portátil e PCA. Souza et al. (2015)[31] realizam análise exploratória a partir das concentrações dos constituintes inorgânicos (sódio, cálcio, magnésio, estrôncio e ferro) do petróleo e PCA. Vieira et al. (2016)[5] aplicam a PCA em dados de densidade API, concentrações de vanádio e níquel para determinar origem de amostras de petróleo em diferentes campos da bacia sedimentar do Espírito Santo.

Antes de iniciar a produção em uma nova jazida petrolífera, são extraídas amostras para análise preliminar (testes de verificação de desempenho e viabilidade) a partir da rocha reservatório. Desta amostra, são obtidas quantidades inferiores a 1 litro de petróleo o que impossibilita uma caracterização físico-química completa do óleo devido ao reduzido volume de amostra. Entretanto, a partir de um banco de dados com características desconhecidas de poços produtores pode ser possível estimar as propriedades físico-químicas de um novo reservatório ou relacionar este novo reservatório com algum ja existente.

2. Objetivo

Esse artigo objetiva identificar perfis de petróleo utilizando a técnica de Espectroscopia na região do Infravermelho Médio (MIR) associada a técnicas quimiométricas de reconhecimento de padrões. A PCA foi utilizada para a redução da dimensão da matriz de dados de infravermelho seguido pela aplicação do k-NN para a identificar amostras conhecidas mais similares à um conjunto de amostras desconhecidas.

3. Metodologia

Neste trabalho foram utilizados 81 petróleos, no qual foram adquiridos espectros MIR e medidas 13 propriedades físico-químicas. Estes dados formaram duas matrizes: uma contendo os espectros MIR das amostras outra contendo os resultados dos ensaios físico-químicos. Visando identificar amostras mais similares, cada conjunto de dados foi, separadamente, submetido à PCA seguida pela k-NN. As 81 amostras foram divididas em 69 amostras para o conjunto treinamento (construção do modelo) e 12 amostras para o conjunto de teste (verificação da performance do modelo). A PCA foi utilizada para redução da dimensão seguida pela aplicação do kNN visando determinar os 5 vizinhos mais próximos a cada amostra do conjunto teste. Os resultados de cada conjunto de dados foram comparados quanto a similaridade na identificação das amostras próximas.

3.1 Amostras

As 81 amostras de petróleo utilizadas neste trabalho são oriundas da bacia sedimentar da costa brasileira. O procedimento aplicado na transferência das amostras para realização dos ensaios físico-químicos seguiu a norma padrão ASTM D5854.[32] Este método descreve sobre procedimentos como manuseio, mistura e condicionamento requerido para garantir uma amostra representativa de petróleo. Os ensaios físico-químicos foram realizados segundo métodos padronizados ou desenvolvidos por este grupo de pesquisa, conforme mostrado na Tabela F.1. Todos os resultados destes ensaios são mostrados no material suplementar (Tabela E.1, Anexo E).

3.2 Obtenção dos espectros MIR

Os espectros de infravermelho na região do médio foram adquiridos em um espectrômetro do fabricante Perkin Elmer, modelo Spectrum 400 utilizando acessório de célula horizontal de ATR constituída de ZnSe 45º fabricado pela Pike Technologies. Para cada amostra foram coletados 32 *scans*, em triplicata autêntica, com resolução de 4 cm^{-1} e intervalo de leitura de 4000-650 cm^{-1}. O tempo para analisar as triplicatas foi em torno de 10 minutos sendo necessário aproximadamente 1 mL de amostra por ensaio.

3.3 Tratamento quimiométrico

Os dados foram submetidos a PCA para a redução da dimensão seguida pela aplicação do kNN para a identificação das amostras mais similares de petróleo, ou seja, proximidade amostral no subespaço.

3.3.1 Análise por Componentes Principais – PCA

A PCA é aplicada na transformação de grandes matrizes de dados em poucas variáveis, denominadas componentes principais (h), as quais retém a maior parte da variância (informação). Esta transformação tem por objetivo de separar a informação que descreve a estrutura dos dados daquela que não descreve (ruído). Uma matriz de dados $\mathbf{X}$ contendo n amostras por m variáveis (propriedades físico-químicas ou número de ondas espectrais) pode ser vista como um arranjo de dados com dois modos. O primeiro modo contém informações a respeito das amostras (m), representada pela matriz escores (**T**) enquanto o segundo modo contém informações a respeito das variáveis (n), representada pela matriz peso (**P**). Normalmente variáveis espectrais possuem alto grau de correlação, dada por:

$$\mathbf{X}_{(n,m)} = \mathbf{T}_{(n,h)} \cdot \mathbf{P}^{T}_{(m,h)} + \mathbf{E}_{(n,m)} \qquad \text{(F.3.1)}$$

A redução da dimensionalidade para h componentes principais simplifica a observação do agrupamento ou não de amostras por similaridade, ou seja, a proximidade entre as amostras no espaço. Amostras próximas, no espaço dos principais componentes, tendem a ter características físico-químicas semelhantes. [23,27,36,37] Para a verificação e cálculo dessa proximidade entre as amostras, é essencial usar uma métrica, como é o caso da Euclidiana no método kNN.

3.3.2 K vizinhos mais Próximos – k-NN

A regra dos k vizinhos mais próximos (k-NN)[23,37] é um método para categorizar amostras de teste de acordo com sua proximidade, no espaço, em relação às amostras já conhecidas (amostras do conjunto de treinamento). Na utilização do método, calcula-se a distância das k amostras mais próximas (vizinhas), pois parte-se do princípio de que quanto mais similares forem as amostras, menores serão suas distâncias no espaço. Neste trabalho, o método k-NN foi otimizado para os 5 vizinhos mais próximos. A proximidade entre amostras no espaço foi determinada utilizando a métrica euclidiana:

$$d_{Euclidiana}(A, B) = \left\| \mathbf{x}_A - \mathbf{x}_B \right\| \qquad \text{(F.3.2)}$$

onde $\mathbf{x}_A$ e $\mathbf{x}_B$ são duas amostras quaisquer no espaço das componentes principais.

3.3.3. Construção do modelo quimiométrico

Das 81 amostras, 69 foram separadas para o conjunto de treinamento, utilizadas para construção dos modelos e 12 amostras como conjunto de teste para a verificação da performace do modelo. As amostras de teste foram selecionadas aleatoriamente considerando a faixa de propriedades físico-químicas, assim como as amostras com valores extremos de propriedades foram evitadas. Os modelos de classificação foram contruídos no *software* MATLAB versão 7.12. Para o pré-processamento dos espectros MIR foram aplicados: variação normal padrão (SNV - *Standard Normal Variate*)[38] seguido da centralização na média. Para o conjunto de dados das propriedades físico-químicas (Tabela F.1), os dados foram autoescalonados. A partir dos escores do modelo PCA foi aplicado o kNN. Os resultados da modelagem obtidos via MIR foram comparados com os resultados obtidos a partir das propriedades físico-químicas.

Tabela F.1 Propriedades físico-químicas das amostras de petróleo

Nº	Propriedades físico-químicas	Método de análise	Faixa
P1	Gravidade API	ISO 12185[39]	17 – 46,3
P2	Pressão de vapor Reid (kPa)	ASTM D323[40]	10,06 – 73
P3	Ponto de fluidez (Máximo) (ºC)	ASTM D5853[41]	-5 – 21
P4	Ponto de Névoa (ºC)	DSC (1°C/min)[42]	12,14 – 36,53
P5	Viscosidade cinemática 40ºC ($mm^2.s^{-1}$)	ASTM D7042[43]	3,15 – 305,1
P6	Resíduo de carbono micro (wt%)	ASTM D 4530[44]	0,29 – 8,15
P7	Saturados (massa%)	SFC/TLC-FID[45,46]	36,7 – 82,0
P8	Aromáticos (massa%)	SFC/TLC-FID[45,46]	8,5 – 37,2
P9	Resinas (massa%)	SFC/TLC-FID[45,46]	0,0 – 42,8
P10	Asfaltenos (massa%)	ASTM D6560[47]	0.0 – 4.2
P11	Teor de enxofre (massa%)	ASTM D4294[48]	0,04 – 0,815
P12	Teor de nitrogênio (mg/kg)	ASTM D4629[49]	76 – 4817
P13	Número de acidez total (mgKOH/g)	ASTM D664[50]	0,03 – 3,09

4. Resultados e discussão

Uma amostra de petróleo contém uma grande quantidade de componentes no qual influencia suas características físico-químicas. A técnica de espectroscopia na região do infravermelho médio obtém informação química a nível molecular do petróleo. No entanto, a utilização da PCA associada ao k-NN possibilita o reconhecimento de amostras de petróleo similares, chegando a mesma conclusão se o analista utilizar um banco de resultados das amostras.

As 69 amostras selecionadas para construção do modelo apresentam grande variabilidade na distribuição dos dados para as 13 propriedades físico-químicas (Tabela F.1). A densidade API apresenta uma variação de óleo pesado (17,0) a leve (46,3) abrangendo uma ampla faixa de perfis de petróleo conhecidos da bacia sedimentar da costa brasileira. Os espectros MIR apresentaram pequena variação na linha base (Figura F.1a), na qual foi corrigida pela aplicação do pré-processamento SNV (Figura F.1b). Após aplicação do SNV verifica-se a suavização do espectro reduzindo a influência da variação da linha de base.

A espectroscopia MIR tem importante aplicação na identificação de compostos orgânicos. O petróleo é constituido predominantemente por hidrocarbonetos. Entretanto, devido a sua complexidade é díficil a identificação de bandas específicas em seu espectro. Na Figura F.1, a partir do espectro MIR é possível identificar um perfil próprio do petróleo. A intensa banda na região de 3000-2700 cm^{-1} refere-se ao estiramento C-H. No entanto, essa região pode ainda conter informações sobre grupos ácidos uma vez que a faixa de absorção do OH de ácido é de aproximadamente 3400-3000 cm^{-1}. Na região de impressão digital, entre 1300 e 900 cm^{-1}, existe informações relevantes de possíveis estiramentos de ligações C-O de ácidos, álcoois, fenóis, éteres, ésteres. Na última região, entre 900 e 500 cm^{-1}, há informações de compostos aromáticos, e heteroaromáticos. Estas informações estão presentes nos espectros de cada amostra, no entanto, faz-se necessário a aplicação da quimiometria para extração desta informação que está relacionada com o perfil do óleo, caracteristico de cada reservatório. Nota-se, ainda, que não há absorções significativas na região de 3600-3200 cm^{-1}, relacionadas ao estiramento N-H e O-H da água, mostrando que provavelmente esses constituintes se apresentam em pequenas quantidades nas amostras trabalhadas ou a técnica não apresentou sensibilidade suficiente para a detecção desses grupos nessa região.[16] Desse modo, por análise direta dificilmente poderia se afirmar a presença de eventuais grupos orgânicos, porém com o auxilio da análise multivariada é possível a extração desse tipo de informação para discriminar amostras de petróleo.

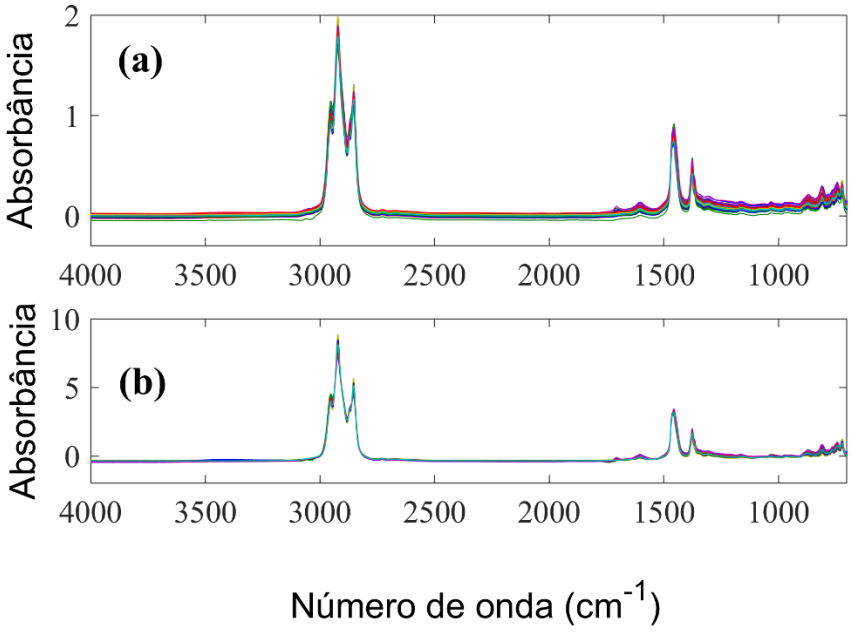

Figura F.1 Espectro MIR de amostras de petróleo (a) original e (b) após pré-processamento com SNV.

Os modelos k-NN foram construídos a partir dos escores da PCA dos dados de treinamento. As variâncias explicadas para as dez primeiras componentes principais dos dois modelos estão mostradas na Figura F.2. O modelo contendo dados espectroscópicos foi construído com as 5 primeiras componentes principais retendo 93,3% da variância dos dados, enquanto que para o conjunto de dados físico-químico foi retido 86,7% da variância (Figura F.2).

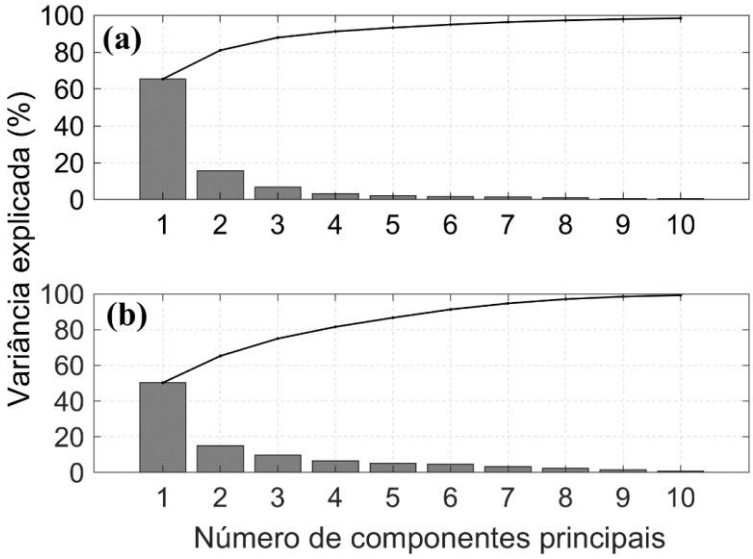

Figura F.2 Variância explicada dos modelos de PCA construídos com dados do espectro infravermelho médio (a) e das propriedades físico-químicas das amostras (b).

Na Figura F.3, é mostrado o gráfico dos escores da PCA com as 69 amostras de treinamento e 12 amostras de teste (t1 a t12). Essa representação gráfica bidimensional foi construída apenas com as duas primeiras componentes mostrando que o perfil das amostras de teste não difere das amostras de treinamento, pois estas apresentam-se como um grande agrupamento de pontos.

Na Figura F.4 é mostrado uma representação em 2D das amostras de treinamento mais próximas de cada amostra de teste. Esta representação está em duas dimensões, no entanto, é possível extrapolar para mais dimensões visando reter maior variância dos dados. Amostras de petróleos com características físico-químicas semelhantes devem apresentar menor distância no gráfico dos scores da PCA.

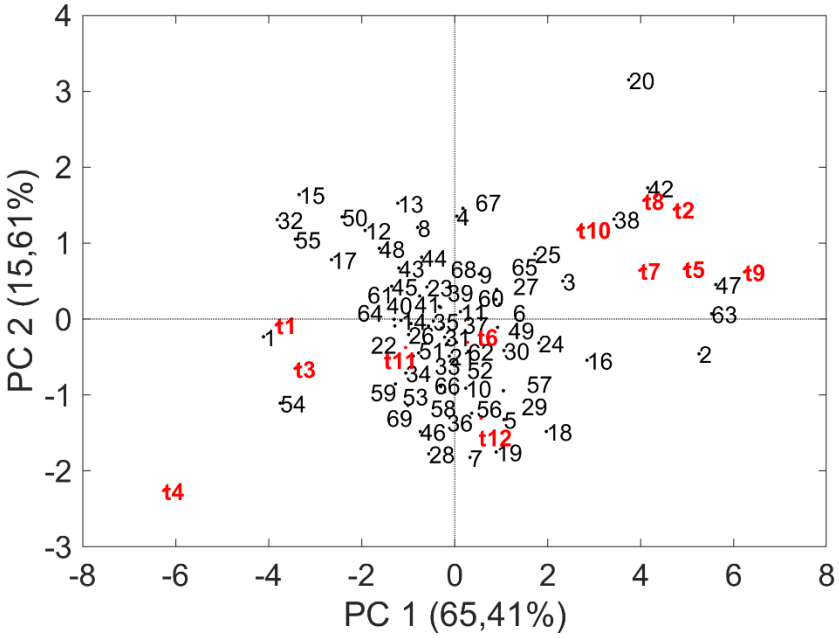

Figura F.3 Gráfico de escores 2D (PC1 e PC2) do PCA usando os espectros MIR das amostras de treinamento.

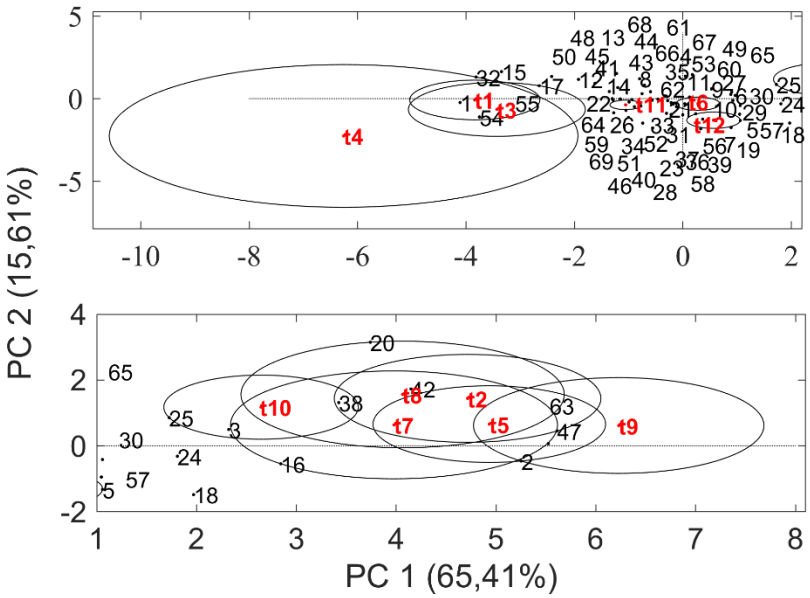

Figura F.4 Gráfico de escores 2D (PC1 e PC2). As amostras de teste são destacadas em negrito: (a) amostras t1, t3, t4, t6, t11 e t12; (b) amostras t2, t5, t7, t8, t9 e t10. As elipses representam as cinco amostras mais próximas.

Identificar amostras com características semelhantes é importante para a previsão da qualidade do óleo na avaliação de petróleo. Para discriminar as amostras mais similares, foi utilizado o método dos vízinhos mais próximos. Estes foram determinados nos dois modelos PCA, o modelo construído com as propriedades físico-químicas e com os espectros MIR. Para a Amostra de teste t1, os cinco vizinhos mais próximos observados com dados MIR foram as amostras: A1, A60, A61, A34, e A19, nesta ordem de proximidade. Destas, apenas as amostras A55 e A17 não estão entre as 5 mais próximas no modelo construído com dados das propriedades físico-químicas. Para a Amostra de teste t2, foram identificadas quatro amostras de treinamento em comum nos dois modelos. Para as Amostras t3, t4, t6, t7, t10, t11 e t12 apenas duas amostras apresentaram diferenças dos resultados entre os modelos, enquanto que para a Amostra t5, t8 e t9 apenas uma amostra não está entre as 5 mais próximas do modelo construído com dados de propriedades físico-químicas (Tabela F.2).

Esses resultados indicam que o modelo construído a partir dos espectros MIR geram informação semelhante aos modelos criados com as propriedades físico-químicas. Em determinadas situações, não há restrita prioridade em se caracterizar uma amostra de petróleo, mas sim, em se ter uma estimativa das propriedades físico-

químicas. Em amostra de teste de verificação de performance (PVT) são extraídos volumes inferiores a 1 litro de petróleo e a partir desta amostra, estima-se a qualidade do óleo e, consequentemente, seu valor comercial. Durante a etapa de destilação do petróleo, ajustes são realizados nas condições do processo baseados no banco de dados de destilação do referido poço. Assim, a utilização da espectroscopia na região do infravermelho médio pode fornecer informações úteis sobre similaridades de petróleos visando prever seu comportamento na destilação ou subprodutos que podem ser extraídos.

Tabela F.2 Amostra de teste e respectivas amostras de treinamento por ordem de proximidade.

Amostra teste	**Amostras treinamento**				
t1	A1	A54	A55*	A32	A17*
t2	A42	A38	A47*	A63	A20
t3	A54*	A1	A17*	A55	A32
t4	A54	A1*	A32	A55	A17*
t5	A47	A63	A2	A42*	A38
t6	A62	A52*	A60	A11*	A21
t7	A38	A42	A16*	A47*	A63
t8	A42	A38	A20	A47*	A3
t9	A47	A63	A2	A42*	A38
t10	A3*	A38	A25*	A42	A24
t11	A26*	A66	A51	A34	A41*
t12	A29	A56*	A5	A10	A19*

* Amostras que não estão entre os vizinhos mais próximos observados no modelo usando os dados de propriedades físico-químicas.

Pelo gráfico de escores e loadings (Figura F.5), é possível verificar quais são as propriedades mais relacionadas a cada amostra teste utilizando as duas primeiras componentes principais. Entretanto, outras relações entre propriedades e amostras podem ser verificadas usando outras componentes. Nesse caso, com as duas primeiras componentes, a Amostra t1 possui relação direta com P1 (grau API) e relação inversa com P11 (Teor de Enxofre). O mesmo pode ser verificado pelo gráfico de correlação (Figura F.6) em que a propriedade P1 apresenta uma relação inversa

com a propriedade P11. Pelos dados da (Tabela E.1, Anexo E), isso pode ser verificado, amostras com grau API, apresentam menor teor de enxofre. Para as amostras t1, t3 e t4 pelo gráfico de escores e pesos (Figura F.5), é possível identificar uma relação direta desse grupo de petróleo com P1 e P7 (saturados) e uma relação indireta com P6 (resíduo de microcarbono), P11 e P12 (teor de nitrogênio). A partir do gráfico de correlação (Figura F.6), temos relações positivas entre essas propriedades (P1 e P7). Percebe-se também que o grupo das amostras t1, t3 e t4 apresenta similaridade entre as amostras de cada grupo, mostrando valores semelhantes das propriedades P1 e P7 (Tabela E.1, Anexo E). Ao comparar a propriedade P1, entre a amostra t1 e mais próximas (Tabela F.2), na (Tabela E.1, Anexo E), verificou-se valores próximos que variam de 35,9 a 46,4 API de gravidade, valores característicos de óleos leves.[1]

Para as amostras t2, t5, t7, t8 e t9 (Figura F.5), há uma relação direta com as propriedades P5 (viscosidade cinemática a 40 ° C) e P13 (número de acidez total). Observa-se que os valores entre as amostras em cada propriedade são próximos e representam as amostras com os valores mais elevados do conjunto amostral (Tabela E.1, Anexo E). Também é importante discutir que essas amostras apresentam altos níveis de compostos aromáticos, mostrando alguma relação com a propriedade P8. A Figura F.6 mostra que essas propriedades têm um grau de correlação positivo entre elas, sendo esse maior grau entre as propriedades P5 e P13. (Tabela E.1, Anexo E). Ao comparar os resultados da amostra t2 e seus vizinhos mais próximos com a propriedade P13, observa-se que são amostras com alto teor de ácido. Essa análise é de interesse industrial, uma vez que petróleos com alto teor de acidez, superior a 0,5 mg $KOH \cdot g^{-1}$, podem causar corrosão no processo de produção.[51] Usando este gráfico (F.5), é possível identificar rapidamente o grupo que possui um alto teor de ácido.

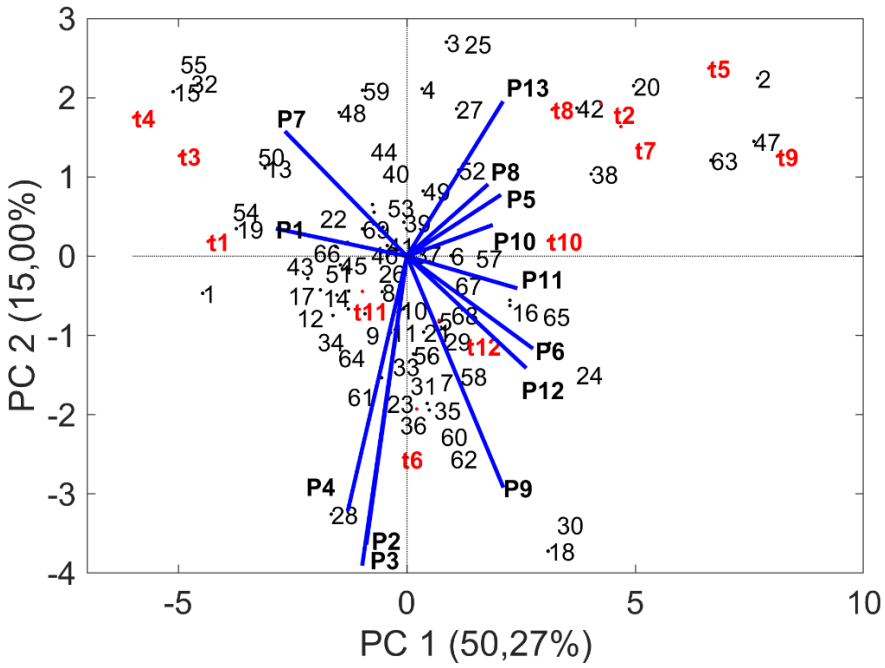

Figura F.5 Biplot de escores e pesos (PC1 e PC2) para as propriedades físico-químicas de amostras de treinamento e teste (t1 a t12).

As amostras t6 e t12 apresentam valores próximos e elevados de resina (P9). O grupo de amostra de teste t10 mostra boa similaridade para as propriedades P5, P8, P10 (asfaltenos) e P11 (Figura F.5) e uma relação inversa com P1. Essa relação entre propriedades também pode ser avaliada pelo gráfico de correlação (Figura F.6). A amostra t11 (Figura F.5) está inversamente relacionada às propriedades P5, P8 e P10, estando elas correlacionadas positivamente (Figura F.6). Assim, essas informações mostram que essas amostras apresentam altos teores de aromáticos e asfaltenos, justificando seus maiores valores de viscosidade.[5,14]

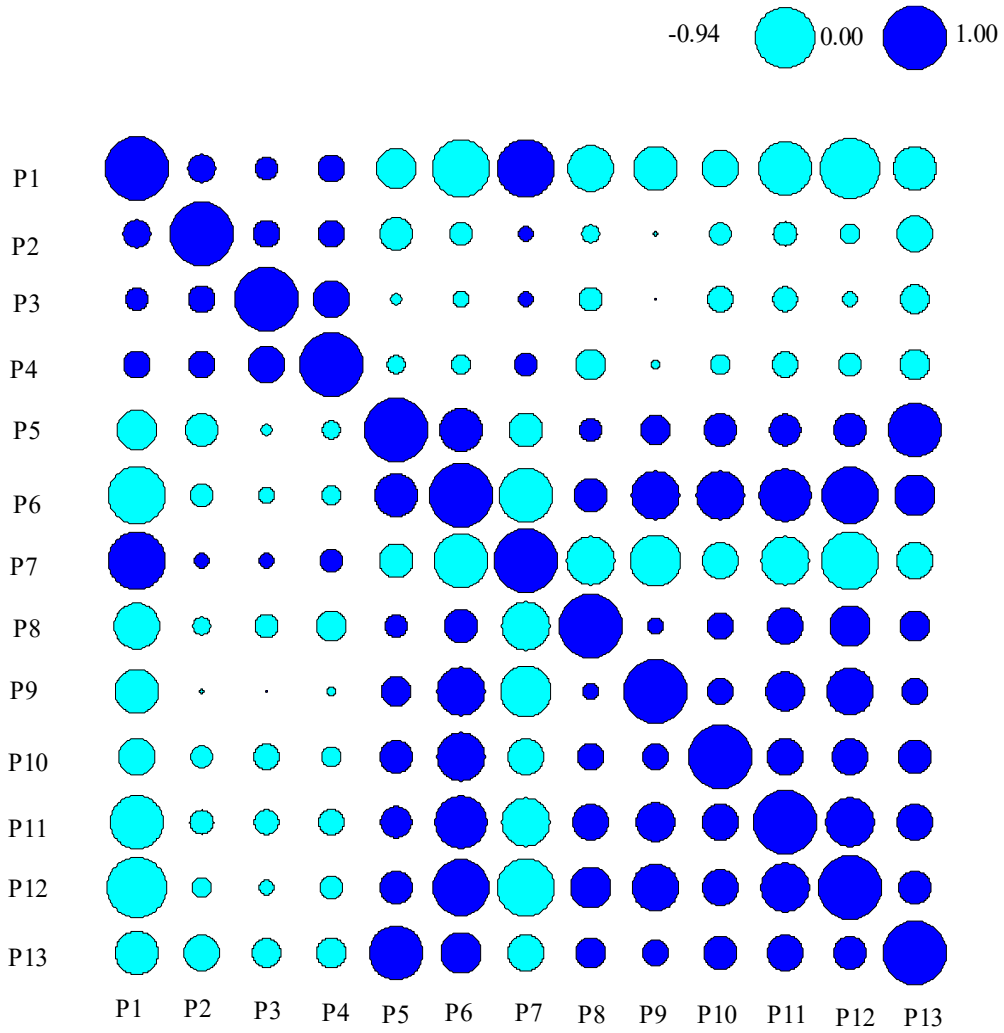

Figura F.6. Gráfico de correlação das propriedades físico-químicas de amostras de petróleo.

A partir dessa análise de correlação entre propriedades e amostras vizinhas utlizando o método do k-NN, método não supervisionado, é possível inserir nesse banco de dados, amostras de petróleo com perfil químico desconhecido e realizar o reconhecimento de padrão por meio de técnicas espectrocópicas. É importante enfatizar que o modelo construído não é universal; no entanto, a metodologia usada pode ser universal e ter aplicação expandida em óleos de outros países, dependendo apenas do histórico de produção e de um banco de dados de espectros.

5.5 Conclusões

A utilização da espectroscopia na região do infravermelho médio associada aos métodos multivariados de reconhecimento de padrões PCA e k-NN constituem uma

ferramenta eficiente na identificação de perfis de petróleo. Esta metodologia possibilita obter conclusões equivalentes sobre óleos mais similares, ou seja, com comportamentos semelhantes, quando são utilizados espectros MIR ou resultados de ensaios físico-químicos dos petróleos. Os resultados nas amostras de teste mostraram que dos cinco vizinhos mais próximos, pelo menos três eram compatíveis com as propriedades espectrais e físico-químicas do modelo k-NN construído. Quatro amostras de teste mostraram 80% de compatibilidade dos vizinhos mais próximos na comparação dos modelos (modelos PCA e k-NN). Em geral, os modelos construídos (propriedades espectrais e físico-químicas) apresentaram 66,7% de compatibilidade. A metodologia desenvolvida viabiliza a previsão da qualidade do óleo na avaliação do petróleo por um método mais rápido e em menor volume de amostra.

5.6 Referências Bibliográficas

1. SPEIGHT J. G. Handbook of Petroleum Product analysis. 3ª ed. New York: Wiley-Interscience; 2002.

2. ANP - Agência Nacional do Petróleo, Gás Natural e Biocombustíveis - Petróleo e derivados. Disponível em: http://www.anp.gov.br. Acessado em 07 de outubro de 2019

3. RIAZI M. R. Characterization and properties of petroleum fractions. 1ª ed. Philadelphia: ASTM Stock Number: MNL50; 2005, 50p.

4. BARBOSA, L. L.; SAD, C. M, S.; MORGAN, V. G.; SANTOS, M. F. P.; CASTRO, E. V. R. Time-Domain Proton Nuclear Magnetic Resonance and Chemometrics for Identification and Classification of Brazilian Petroleum. *Energy Fuels.* **27**,6560-6566, 2013.

5. VIEIRA, L. V. V.; RAINHA, K. P.; CASTRO, E. V. R.; FILGUEIRAS, P. R.; CARNEIRO, M. T. W. D.; BRANDÃO, G. P. Exploratory data analysis using API gravity and V and Ni contents to determine the origins of crude oil samples from petroleum fields in the Espírito Santo Basin (Brazil). Microchemical Journal, **124**, 26-30, 2016.

6. EL-GAYAR, M. S.; MOSTAFA, A. R.; ABDELFATTAH, A. E.; BARAKAT, A. O. Aplication of geochemical parameters for classification of crude oils from Egypt into source-related types. *Fuel Process*. *Technol*., **79**, 13-28, 2002.

7. CURIALE, J. A.; FROLOV, E. B. Occurrence and origin of olefins in crude oils. A critical review. *Org. Geochem*.**1**, 397-408, 1998.

8. SOMBRAL, L. G. S.; PICKLER, A. C.; AIRES, J. R.; RIEHL, C. A. Study on nickel and vanadium removal in thermal conversion of oil sludge and oil shale sludge. *J. Phys. IV France*, **107**, 1255-1258, 2003.

9. BARBOSA L. L.; SAD C. M. S.; MORGAN, V. G.; FIGUEIRAS, P. R.; CASTRO, E. R. V. Application of low field NMR as an alternative technique to quantification of total acid number and sulphur content in petroleum from Brazilian reservoirs. *Fuel*, **176**, 146-152, 2016.

10. ROCHA, J. T. C.; OLIVEIRA, L. M. S.; DIAS, J. C. M.; PINTO, U. B.; MARQUES, M. L. S. P.; OLIVEIRA, B. P.; FILGUEIRAS, P. R.; CASTRO, E. V. R.; OLIVEIRA, M. A. L. Sulfur Determination in Brazilian Petroleum Fractions by Midinfrared and Near-infrared Spectroscopy and Partial Least Squares Associated with Variable Selection Methods. *Energy Fuels*. **30**, 698–705, 2016.

11. DUARTE, L. M.; FILGUEIRAS, P. R.; DIAS, J. C. M.; OLIVEIRA, L. M. S.; CASTRO, E. V. R.; OLIVEIRA, M. A. L. Study of Distillation Temperature Curves from Brazilian Crude Oil by ^{1}H Nuclear Magnetic Resonance Spectroscopy in Association with Partial Least Squares Regression. *Energy Fuels*, **31**, 3892−3897, 2017.

12. SIMANZHENKOV, V.; IDEM R. Crude oil chemistry. 1ª ed. New York: Marcel Dekker, Inc., 2003.

13. LYONS, W. C.; PLISGA, G. J. Standard Handbook of Petroleum & Natural Gas Engeneering. 2ª ed. Amsterdam: Elsevier; 2005.

14. SAD, C. M. S.; LACERDA, J. R. V.; FILGUEIRAS, P. R.; RIGONI, V. S.; BASSANE, J. F. P.; CASTRO, E. V. R.; PEREIRA, K. S.; SANTOS, M. F. P. Limitations of the pour point measurement and the influence of the oil composition on its detection using principal component analysis, *Energy Fuels*, **28**, 1686–1691, 2014.

15. SOYEMI, O. O.; BUSCH, M. A.; BUSCH, K. W. Multivariate analysis of near-infrared spectra using the G-programming language. *J Chem Inf Comput Sci*., **40**,1093-1100, 2000.

16. FILGUEIRAS, P. R.; SAD, C. M. S.; LOUREIRO, A. R.; SANTOS, M. F. P.; CASTRO, E. V. R.; DIAS, J. C. M.; POPPI, R. J. Determination of API gravity, kinematic viscosity and water content in petroleum by ATR-FTIR spectroscopy and multivariate calibration. *Fuel*, **116**, 123-130, 2014.

17. KHANMOHAMMADI, M.; GARMARUDI, A. B.; GUARDIA, M. Characterization of petroleum-based products by infrared spectroscopy and chemometrics. *TrAC Trends Anal Chem*., **35**, 135-49, 2012.

18. PASQUINI, C.; BUENO, A. F. Characterization of petroleum using near-infrared specstrocopy: quantitative modeling for the true bolling ponit curve and specific gravity. *Fuel*, **86**, 1927-1934, 2007.

19. FALLA, F. S.; LARINI, C.; LE ROUX, G. A. C.; QUINA, F. H.; MORO, L. F. L.; NASCIMENTO, C. A. O. Characterization of crude petroleum by NIR. *J. Pet. Sci. Eng*., **51**, 127-137, 2006.

20. KALLEVIK, H.; KVALHEIM, O. M.; SJÖBLOM, J. Quantitative determination of saphaltenes and resins in solution by means of near-infrared spectroscopy. Correlations to emulstion stability. *J Colloid Interface Sci*., **225**, 494-504, 2000.

21. KUPTSOV, A. K.; ARBUZOVA, T. V. A study of heavy oil fractions by fourier-transform near-infrared raman spectroscopy. *Petroleum Chemistry*, **51**, 203-211, 2011.

22. HANNISDAL, A.; HEMMINGSEN, P. V.; SJÖBLOM, J. Group-type analysis of heavy crude oils using vibration spectroscopy in combinations with multivariate analysis*. Ind. Eng. Chem. Res*., **44**, 1349-1357, 2005.

23. BRERETON R. G. Chemometrics: Data Analysis for the laboratory and Chemical Plant. Chichester: John Wiley & Sons Ltda, 2003, 150p.

24. CORREIA, R. M.; TOSATO, F.; DOMINGOS, E.; RODRIGUES, R. R. T.; AQUINO, L. F. M.; FILGUEIRAS, P. R.; LACERDA, V.; ROMÃO, W. Portable near infrared spectroscopy applied to quality control of Brazilian coffee. *Talanta,* **176**, 59-68, 2018.

25. SEKULIC, S.; SEASHOLTZ, M. B.; WANG, Z, KOWALSKI, B. R. Nonlinear multivariate calibration methods in analytical chemistry. *Anal. Chem*., **19**, 835-845, 1993.

26. WOLD, S.; SJOSTROM, M. Chemometrics, present and future success. *Chemom. Intell. Lab. Syst*., **44**, 3-14, 1998.

27. BRO, R.; SMILDE, A. K. Principal component analysis. *Anal. Methods*, **6**, 2812-2831, 2014.

28. URDAL K, VOGT NB, SPORSTØL SP, LICHTENTHALER RG, MOSTAD H, KOLSET K, NORDENSON S, ESBENSEN K. Classification of weathered crude oils using multimethod chemical analysis, statistical methods and SIMCA pattern recognition. *Mar. Pollut. Bull*, **8**, 366-373, 1986.

29. CHIABERGE, S.; FIORANI, T.; SAVOINI, A.; BIONDA, A.; RAMELLO, S.; PASTORI, M.; CESTI, P. Classification of crude oil samples through statistical analysis of APPI FTICR mass spectra. *Fuel Process. Technol*. ***106***,181-185, 2013.

30. ALTINPINAR, S.; SORAK, D.; SIESLER, H. W. Near infrared spectroscopic analysis of hydrocarbon contaminants in soil with a hand-held spectrometer. *J. Near Infrared Spectrosc*., **21**, 511-521, 2013.

31. SOUZA, M. O.; RAINHA, K. P.; CASTRO, E. V. R.; CARNEIRO, M. T. W. D.; FERREIRA, R. Q. Análise exploratória das concentrações dos metais Na, Ca, Mg, Sr e Fe em extrato aquoso de petróleo, determinados por ICP OES, após otimização empregando planejamento de experimentos. *Quim. Nova*, **7**, 980-986, 2015.

32. ASTM. Standard test method for mixing and handling of liquid of petroleum and petroleum products, D5854. West Conshohocken, PA: ASTM International; 2005.

33. WOLD S. Principal component analysis. *Chemom. Intell. Lab. Syst*., **2**, 37-52, 1987.

34. PEARSON K. On lines and planes of closest fit to systems of points in space. *Philos. Mag*., **2**, 559-572, 1901.

35. HOTELLING H. J. Analysis of a complex of statistical variables into principal components. *Educ. Psych*., **24**, 417-441, 1933.

36. ABDI, H.; WILLIAMS, L. J. Principal component analysis. Wiley *Interdisciplinary Reviews Computational Statistics*, **2**, 433-459, 2010.

37. TIPPING, M. E.; BISHOP, C. M. Probabilistic principal component analysis. *J. R. Statist. Soc. B*., **61**, 611-622, 1999.

38. Fearn, T.; Riccioli, C.; Garrido-Varo, A.; Guerrero-Ginel, J. E. On the geometry of SNV and MSC. *Chemom. Intell. Lab. Syst*. **1**, 22-26, 2009.

39. ISO. Determination of density - oscillating u-tube method, 12185. Geneva, SZ: International Organization for Standardization; 1996.

40. ASTM. Standard test method for vapor pressure of petroleum products (Reid Method), D323. West Conshohocken, PA: ASTM International; 1968.

41. ASTM. Standard test method for pour point of crude oils, D5853. West Conshohocken, PA: ASTM International; 2009.

42. ASTM. Standard practice for viscosity-temperature charts for liquid petroleum products, D341. West Conshohocken, PA: ASTM International; 2009.

43. ASTM. Standard test method for dynamic viscosity and density of liquids by stabinger viscometer (and the calculation of kinematic viscosity), D7042. West Conshohocken, PA: ASTM International; 2007.

44. ASTM. Standard test method for determination of carbon residue (micro method), D4530. West Conshohocken, PA: ASTM International; 2003.

45. ASTM D2549-02 Standard Test Method for Separation of Representative Aromatics and Nonaromatics Fractions of High-Boiling Oils by Elution Chromatography. Annual Book of ASTM Standards; American Society for Testing Materials: Philadelphia, PA, 1989.

46. FILGUEIRAS, P. R.; PORTELA, N. A.; SILVA, S. R. C.; CASTRO, E. V. R.; OLIVEIRA, L. M. S. L.; DIAS, J. C. M.; NETO, A. C.; ROMÃO, W. POPPI, R. J. Determination of Saturates, Aromatics, and Polars in Crude Oil by ^{13}C NMR and Support Vector Regression with Variable Selection by Genetic Algorithm. *Energy Fuels*, 30, 1972−1978, 2016.

47. ASTM. Standard test method for determination of asphaltenes (heptane insolubles) in crude petroleum and petroleum products, D6550. West Conshohocken, PA: ASTM International; 2010.

48. ASTM. Standard test method for sulfur in petroleum and petroleum products by energy dispersive x-ray fluorescence spectrometry, D4294. West Conshohocken, PA: ASTM International; 2008.

49. ASTM. Standard test method for trace nitrogen in liquid petroleum hydrocarbons by syringe/inlet oxidative combustion and chemiluminescence detection, D4629. West Conshohocken, PA: ASTM International; 2008.

50. ASTM. Standard test method for acid number of petroleum products by potentiometric titration, D664. West Conshohocken, PA: ASTM International; 2007.

51. BARROS, E. V.; DIAS, H. P; GOMES, A. O.; RODRIGUES, R. R. T.; MOURA, R. R.; SAD, C. M. S.; FREITAS, J. C. C.; NETO, A. C.; AQUIJE, G. M. F. V.; ROMÃO, W. Study of degradation of acid crude oil by high resolution analytical techniques. J. Pet. Sci. Eng., **154**,194-203, 2017.