UNIVERSIDADE SANTA CECÍLIA
PROGRAMA DE PÓS-GRADUAÇÂO EM ENGENHARIA MECÂNICA
MESTRADO EM ENGENHARIA MECÂNICA

SAMUEL PINHEIRO GONÇALVES

APRENDIZADO DE MÁQUINA APLICADO NA CLASSIFICAÇÃO DE TIPOS DE ÓLEOS LUBRIFICANTES DE MOTORES AUTOMOTIVOS A PARTIR DE DADOS DE ESPECTROSCOPIA RAMAN

SANTOS/SP
2024

SAMUEL PINHEIRO GONÇALVES

APRENDIZADO DE MÁQUINA APLICADO NA CLASSIFICAÇÃO DE TIPOS DE ÓLEOS LUBRIFICANTES DE MOTORES AUTOMOTIVOS A PARTIR DE DADOS DE ESPECTROSCOPIA RAMAN

Dissertação apresentada a Universidade Santa Cecília como parte dos requisitos para obtenção de título de mestre no Programa de Pós-Graduação em Engenharia Mecânica, sob a orientação do Prof. Dr. João Inácio da Silva Filho e coorientação do Prof. Dr. Landulfo Silveira Jr.

SANTOS/SP

2024

629.255 Gonçalves, Samuel Pinheiro.
G629a Aprendizado de máquina aplicado na classificação de tipos de óleos lubrificantes de motores automotivos a partir de dados de espectroscopia de Raman /Samuel Pinheiro Gonçalves.
2024.
65 f.

Orientador: Dr. João Inácio da Silva Filho.
Coorientador: Dr. Landulfo Silveira Jr.

Dissertação (Mestrado) - Universidade Santa Cecília, Programa de pós-graduação em Engenheira Mecânica, Santos, SP, 2024.

1. Aprendizado de Máquinas. 2. Espectroscopia Raman. 3. Classificação de óleos lubrificantes. 4. Avaliação de desempenho. l. Silva Filho, João Inácio da. II. Aprendizado de máquina aplicado na classificação de tipos de óleos lubrificantes de motores automotivos a partir de dados de espectroscopia de Raman.

Elaborado pelo SIBi - Sistema Integrado de Bibliotecas - Unisanta

DEDICATÓRIA

Dedico este trabalho a meu Deus, a minha mãe por todo incentivo, a minha esposa e minhas filhas que me apoiaram de diversas maneiras durante esta importante etapa de minha vida.

AGRADECIMENTOS

Agradeço sinceramente a todos os envolvidos na elaboração desta dissertação. Primeiramente, gostaria de expressar minha gratidão aos professores e orientadores, cujo apoio, conhecimento e dedicação foram fundamentais para o desenvolvimento deste trabalho. À instituição, pelos recursos e ambiente propícios à pesquisa, agradeço por disponibilizar as ferramentas necessárias para alcançar os objetivos propostos nesta dissertação.

Agradeço também ao corpo docente e a todo o pessoal de apoio, cujo comprometimento e profissionalismo contribuíram para o êxito desta dissertação. Sem o seu apoio, essa conquista não seria possível.

Por fim, quero estender meus agradecimentos a todos os colegas, amigos e familiares que me apoiaram e encorajaram durante todo esse percurso acadêmico. Seus estímulos foram essenciais para superar os desafios e alcançar os resultados apresentados nesta dissertação.

Obrigado a todos que, de uma forma ou outra, contribuíram para a conclusão bem-sucedida deste trabalho. Sinto-me imensamente grato por poder contar com uma rede de apoio tão competente e dedicada. Que este trabalho possa trazer contribuições significativas à área de pesquisa e ao conhecimento em geral.

RESUMO

Técnicas de Inteligência Artificial (IA) que incluem Aprendizado de Máquina (AM) têm sido utilizadas com sucesso no apoio à classificação de insumos, aditivos e produtos finais para obtenção de informações que reflitam qualidade na tomada de decisão em processos industriais. Esta dissertação visa avaliar três modelos de AM supervisionados, incluindo Floresta Aleatória (*Random Forest*), Máquina de Vetores de Suporte (SVM - *Support Vector Machine*) e Redes Neurais (*Neural Network*), na classificação de tipos de óleos lubrificantes de motores automotivos a partir de dados obtidos por espectroscopia Raman. Os objetivos específicos incluem a escolha e preparação de um conjunto de dados representativos de espectroscopia Raman de óleos lubrificantes, a implementação e ajuste dos modelos de RF, SVM e RN para a classificação dos diferentes tipos de óleos lubrificantes, a avaliação do desempenho dos modelos usando métricas apropriadas e a comparação entre os modelos para identificar seus pontos fortes e fracos. Por fim, esta dissertação também busca fornecer uma discussão sobre os resultados obtidos, destacando a viabilidade do uso de técnicas de AM para a classificação de óleos lubrificantes de motores automotivos baseados em espectroscopia Raman. Os dados foram fornecidos pelo Laboratório de Espectroscopia Vibracional da Universidade Anhembi Morumbi. Foram utilizados três tipos de óleos lubrificantes automotivos: mineral, semissintético e sintético. Os espectros Raman foram coletados e pré-processados para remover ruídos, normalizar os dados e redução de dimensionalidade (PCA). Os dados foram divididos em conjuntos de treinamento e teste usando validação cruzada e métricas de avaliação, como precisão, sensibilidade e F1-*score*, foram utilizadas para avaliar o desempenho dos modelos. Para o óleo lubrificante **mineral**, os modelos RF, SVM e RN apresentaram resultados expressivos em diversas métricas. O RF demonstrou alta precisão (90,9%), enquanto o SVM se destacou pelo melhor recall (94,4%), indicando sua capacidade superior em classificar esses óleos. A análise do F1 score revelou harmonia entre precisão e *recall*, com o SVM liderando (90,7%). Embora todos os modelos apresentassem acurácia favorável, o SVM se destacou (93,5%). No entanto, o modelo RF se sobressaiu com especificidade (95,8%), evidenciando sua habilidade em distinguir óleos não minerais. Para o óleo **semissintético**, observou-se que o SVM alcançou os resultados mais altos em métricas como precisão (93,9%), F1 score (89,9%) e acurácia (93,5%), destacando sua capacidade de evitar falsos positivos e identificar corretamente as instâncias positivas. O RN se destacou com o maior *recall* (88,9%), mostrando um excelente equilíbrio entre precisão e *recall*. No contexto do óleo **sintético**, os modelos SVM e RN se destacaram com resultados excepcionais (100%) em todas as métricas, incluindo precisão, *recall*, F1 score, acurácia e especificidade.

Palavras Chave: Aprendizado de Máquinas. Espectroscopia Raman. Classificação de óleos lubrificantes. Avaliação de desempenho.

ABSTRACT

Artificial Intelligence (AI) techniques that include Machine Learning (ML) have been successfully used to support the classification of inputs, additives and final products to obtain information that reflects quality in decision making in industrial processes. This dissertation aims to evaluate three supervised AM models, including Random Forest, Support Vector Machine (SVM) and Neural Networks, in the classification of types of lubricating oils of automotive engines from data obtained by Raman spectroscopy. Specific objectives include the choice and preparation of a representative Raman spectroscopy data set of lubricating oils, the implementation and tuning of the RF, SVM and RN models for the classification of different types of lubricating oils, the evaluation of the performance of the models using appropriate metrics and comparison between models to identify their strengths and weaknesses. Finally, this dissertation also seeks to provide a discussion on the results obtained, highlighting the feasibility of using AM techniques for the classification of automotive engine lubricating oils based on Raman spectroscopy. The data were provided by the Vibrational Spectroscopy Laboratory at Anhembi Morumbi University. Three types of automotive lubricating oils were used: mineral, semi-synthetic and synthetic. Raman spectra were collected and preprocessed to remove noise, data normalization, and dimensionality reduction (PCA). The data was divided into training and testing sets using cross-validation and evaluation metrics such as accuracy, sensitivity and F1-score were used to evaluate the performance of the models. For mineral lubricating oil, the RF, SVM and RN models presented impressive results in several metrics. The RF demonstrated high precision (90.9%), while the SVM stood out for its best recall (94.4%), indicating its superior ability to classify these oils. The F1 score analysis revealed harmony between precision and recall, with SVM leading (90.7%). Although all models showed favorable accuracy, SVM stood out (93.5%). However, the RF model stood out with specificity (95.8%), demonstrating its ability to distinguish non-mineral oils. For semi-synthetic oil, it was observed that SVM achieved the highest results in metrics such as precision (93.9%), F1 score (89.9%) and accuracy (93.5%), highlighting its ability to avoid false positives and correctly identify positive instances. The RN stood out with the highest recall (88.9%), showing an excellent balance between precision and recall. In the context of synthetic oil, the SVM and RN models stood out with exceptional results (100%) in all metrics, including precision, recall, F1 score, accuracy and specificity.

Keywords: Machine Learning. Raman spectroscopy. Lubricant oil classification. Performance evaluation.

.

LISTA DE FIGURAS

LISTA DE ABREVIATURAS E SIGLAS

AM	Aprendizado de máquina
AP	Aprendizado Profundo
AS	Aprendizado Supervisionado
FN	Falso Negativo
FP	Falso Positivo
IA	Inteligência Artificial
MIR	Espectroscopia no Infravermelho Médio
NIR	Near-Infrared Spectroscopy
NN	Neural Network
RF	Random Forest
RN	Rede Neural
SAE	Society of Automotive Engineers
SVM	Máquina de Vetores de Aprendizado
VN	Verdadeiro Negativo
VP	Verdadeiro Positivo

SUMÁRIO

1 INTRODUÇÃO

1.1 Importância

A aplicação do Aprendizado de Máquina (AM) na classificação de tipos de óleos lubrificantes de motores automotivos a partir de dados de espectroscopia Raman tem sido recentemente estudado como pesquisa científica, como destacado em estudos como o de Passoni, Pacheco e Silveira (2020) e o de Bezerra *et al.* (2021). Com o avanço da tecnologia, a coleta e análise de dados representados por sinais espectroscópicos se tornaram mais acessíveis, possibilitando a utilização de algoritmos de AM para realizar tarefas complexas. Como a identificação e classificação de materiais com base em suas propriedades espectroscópicas, como discutido por Lei *et al.* (2020), onde os autores apresentam em uma revisão abrangente o roteiro para a aplicação do AM no diagnóstico de falhas em equipamentos industriais.

Na indústria automobilística, a precisão na identificação e classificação dos tipos de óleos lubrificantes é necessária para evitar danos aos motores e garantir o correto funcionamento dos veículos (PASSONI; PACHECO; SILVEIRA, 2020). A necessidade do monitoramento de óleos lubrificantes motivou estudos como o de Caneca *et al.* (2006) onde foram empregadas técnicas de espectroscopia no infravermelho médio (MIR) para monitorar as condições de serviço de óleos lubrificantes de motores a diesel. Isso possibilitou obter informações detalhadas sobre a composição molecular dos óleos e diferenciá-los a partir de suas próprias características espectroscópicas.

A análise de óleos lubrificantes permite ainda analisar a degradação e envelhecimento desses materiais, fornecendo informações importantes sobre os mecanismos de desgaste e os processos químicos envolvidos. Passoni, Pacheco e Silveira (2020) e Bezerra *et al.* (2021) empregaram espectroscopia Raman para identificar diferenças na composição de óleos lubrificantes relacionadas a especificações da SAE e aditivos. Pesquisas como a de Yan *et al.* (2022) associaram a espectroscopia Raman com o AM, na classificação de tipos de papel artesanal. Enquanto no estudo de Han *et al.* (2022) foram usadas técnicas de AM para detecção de poluição em óleos lubrificantes agrícolas por meio de espectroscopia de infravermelho próximo (NIR) com o algoritmo *Random Frog.* Esses estudos demonstram a aplicação abrangente de técnicas de espectroscopia em associação com

o AM unindo a perspectiva prática e científica.

Outro aspecto importante a ser considerado é a possibilidade de economia de recursos e redução do impacto ambiental proporcionada pelo uso do AM (LEI *et al.*, 2020), e conforme abordado por Passoni, Pacheco e Silveira (2020) e Bezerra *et al.* (2021) na classificação de óleos lubrificantes. Nesses dois últimos estudos foram exploradas técnicas de análise multivariada buscando a identificação das diferenças na composição de óleos lubrificantes automotivos, conforme especificações da SAE e aditivos. Portanto, conforme esses estudos, verifica-se que com a correta identificação e classificação dos tipos de óleos, é possível otimizar o uso desses materiais, evitando desperdício e reduzindo a necessidade de descarte prematuro.

Considerando as possibilidades de aplicação, bem como suas vantagens, verifica-se a importância de aprofundar as pesquisas de técnicas de Inteligência Artificial (IA) que utilizam o AM na classificação de tipos de óleos lubrificantes de motores automotivos associado com dados de espectroscopia Raman. Com a capacidade de automatizar e agilizar o processo de classificação, além de contribuir de forma significativa para a indústria e a pesquisa científica, essa abordagem pode trazer contribuições importantes na área de lubrificação automotiva (PASSONI; PACHECO; SILVEIRA, 2020; BEZERRA *et al.*, 2021). Dessa forma, destaca-se a necessidade de pesquisa contínua e desenvolvimento de novas técnicas, para aprimorar e elaborar processos de classificação de óleos lubrificantes. Isto contribuirá para ampliar o uso dessa tecnologia, contribuindo para a eficiência, a segurança e a sustentabilidade dos motores automotivos.

1.2 Problemática

Nos últimos anos, observa-se um crescente avanço tecnológico com o uso da IA. Isto trouxe melhorias em diversos setores, destacando-se aqui a indústria automotiva. Porém, a classificação dos tipos de óleos lubrificantes utilizados nos motores automotivos, especialmente no que diz respeito ao tempo de uso, continua sendo um dos principais desafios enfrentados pela indústria (BEZERRA *et al.*, 2021).

De acordo com Xu *et al.* (2023), a aplicação de técnicas de AM na classificação de óleos lubrificantes pode ser feita com resultados satisfatórios. Seu estudo propôs a classificação de tipos de óleo lubrificante através da espectroscopia MIR associada ao algoritmo de análise discriminante linear (LDA - *Linear Discriminant-Analysis*) e

máquina de vetores de suporte (SVM - *Support Vector Machine*), demonstrando eficácia e precisão na classificação desses óleos com base em suas características espectrais.

A problemática reside no fato de que classificar manualmente os óleos lubrificantes é um processo demorado e suscetível a erros (PASSONI, 2017). Soma-se a isso a crescente variedade de óleos disponíveis no mercado, tornando a identificação dos tipos de óleos a partir de suas características físicas e químicas uma tarefa complexa. Estudos recentes, como os realizados por Bezerra *et al.* (2021) usando espectroscopia Raman para identificação e avaliação das alterações químicas decorrentes da temperatura em óleos lubrificantes automotivos oferece informações valiosas para a avaliação do seu estado de degradação.

Neste sentido, a espectroscopia Raman é uma técnica analítica que permite a identificação e análise de substâncias com base nos espectros de dispersão da luz (PASSONI; PACHECO; SILVEIRA, 2020). No entanto, a interpretação manual desses dados pode ser complexa e sujeita a falhas, devido à presença de informações em múltiplas dimensões (FERRARI; BASKO, 2013), que podem estar sobrepostas e gerar interpretações ambíguas. A aplicação das técnicas de AM, como destacado por Peng *et al.* (2022), oferece uma alternativa para trabalhar a complexidade e as múltiplas dimensões de dados.

Assim, o AM associado aos algoritmos de Aprendizagem Supervisionada (AS), surge como uma alternativa viável para a problemática da classificação de tipos de óleos lubrificantes a partir de dados de espectroscopia Raman. O uso de algoritmos e modelos de AM (QURESHI *et al.*, 2023) permite extrair padrões e características dos dados espectroscópicos, treinar um modelo e utilizá-lo para classificar novos óleos com base em suas propriedades espectrais e diferenças espectrais nos grupos (CANECA *et al.*, 2006; PASSONI; PACHECO; SILVEIRA, 2020; BEZERRA *et al.*, 2021), aumentando a eficiência e precisão da classificação.

1.3 Fundamentação teórica

Nos últimos anos, a Aprendizagem de Máquina possibilitou avanços importantes em diversas áreas do conhecimento. O campo do AM experimenta um rápido crescimento e uma aplicação generalizada, conforme abordado por Jordan e Mitchell (2015), que discutem as tendências, perspectivas e prospectivas nessa área. De forma mais direcionada, o trabalho de Kuppusamy, Nikolovski e Teekaraman (2023)

foca sua atenção à aplicação de técnicas de AM para avaliação do desempenho de qualidade de energia em sistemas conectados à rede. Além disso, Lei *et al.* (2020) oferecem uma revisão abrangente e um roteiro para o diagnóstico de falhas em máquinas por meio do AM, destacando a evolução desde os métodos tradicionais até o impacto das teorias de Aprendizado Profundo (AP).

Uma fonte fundamental de dados para análises posteriores é obtida através do espalhamento Raman, em que a luz espalhada por uma amostra manifesta variações em sua frequência decorrentes da interação da luz incidente com a vibração molecular, proporcionando informações detalhadas sobre a estrutura molecular das substâncias em análise. Além disso, é importante conhecer as características do espectro Raman, como a posição dos picos, intensidade e forma das bandas espectrais, que refletem as propriedades moleculares das substâncias analisadas (FERRARI; BASKO, 2013).

Segundo Passoni (2017), a espectroscopia Raman é uma técnica analítica que utiliza a interação entre a luz e a matéria para fornecer informações sobre as moléculas presentes em uma amostra. Isso permite a identificação e caracterização de diferentes componentes presentes nos óleos lubrificantes. Dentre suas aplicações destaca-se o estudo de Passoni, Pacheco e Silveira (2020), que empregou a espectroscopia Raman para identificar diferenças na composição de óleos lubrificantes automotivos e classificá-los, de acordo às especificações de viscosidade e tipo de óleo, a saber: mineral, semissintético, sintético.

O trabalho de Bezerra *et al.* (2021) abordou a identificação e caracterização das alterações químicas e moleculares induzidas pela temperatura em óleos lubrificantes automotivos, utilizando a espectroscopia Raman e técnica multivariada como ferramenta para avaliação. Dá-se ênfase às mudanças químicas sofridas pelos óleos lubrificantes quando submetidos a diferentes condições de temperatura. Com isso, é possível compreender a influência do calor nas propriedades dos lubrificantes, como degradação (ligações duplas em alcenos) e formação de compostos (carbono amorfo).

Apesar da qualidade informativa do método analítico, sua interpretação é manual e apresenta desafios consideráveis, sendo suscetível a erros devido à complexidade dos dados em múltiplas dimensões (PASSONI, 2017). Para contornar essas dificuldades, é necessário um pré-processamento para tratar e normalizar os dados gerados, bem como reduzir sua dimensionalidade para aumentar a eficiência computacional.

Por isso, a literatura destaca a importância da redução de dimensionalidade na análise de dados complexos. Artigos como "*Principal Component Analysis: A Review and Recent Developments*" (JOLLIFFE; CADIMA, 2016) oferecem informações importantes sobre a técnica de análise de componente principal (PCA – *principal component analysis*), uma técnica eficaz para simplificar e interpretar grandes conjuntos de dados, como os provenientes da espectroscopia Raman.

A implementação da PCA não apenas otimiza a eficiência computacional, mas também permite uma identificação precisa dos diferentes tipos de óleos lubrificantes, trazendo avanços na manutenção da qualidade automotiva. Conforme artigos supracitados, como "*Raman Spectroscopy for the Identification of Differences in the Composition of Automobile Lubricant Oils Related to SAE Specifications and Additives*" de (PASSONI; PACHECO; SILVEIRA, 2020) e "*Morphology, Structure and Chemistry of Extracted Diesel Soot-Part I: Transmission Electron Microscopy, Raman Spectroscopy, X-ray Photoelectron Spectroscopy and Synchrotron X-ray Diffraction Study*" de (PATEL *et al.*, 2012), fortalecem a abordagem proposta ao destacar a importância do tratamento e redução de dimensionalidade usando a PCA para a eficácia na análise de dados complexos como os espectros Raman.

Já no âmbito do AM, torna-se necessária a compreensão dos diferentes tipos de algoritmos de Aprendizado Supervisionado (AS) utilizados na classificação, tais como Floresta Aleatória (RF - *Random Forest*), Máquina de Vetores de Suporte (SVM – *Support Vector Machines*), Redes Neurais (NN - *Neural Network*), entre outros modelos (BREIMAN, 2001; CHERKASSKY; MA, 2004; KUPPUSAMY; NIKOLOVSKI; TEEKARAMAN, 2023; QURESHI *et al.*, 2023). Também é útil compreender as etapas do processo de aprendizagem da máquina, como seleção e preparação dos dados, treinamento do modelo e avaliação de desempenho (HE; GARCIA, 2009; MARRS; NI-MEISTER, 2019).

Nesta pesquisa foram utilizadas na classificação três tipos de técnicas de aprendizado de Máquina:

1. Floresta Aleatória (RF - *Random Forest*),
2. Máquina de Vetores de Suporte (SVM – *Support Vector Machines*) *e,*
3. Redes Neurais (NN - *Neural Network*).

1.3.1 Modelos de AM

Os três modelo de AM utilizados neste trabalho serão descritos a seguir:

1.3.1.1 Floresta Aleatória (RF - *Random Forest*)

O primeiro modelo denominado Floresta Aleatória (RF - *Random Forest*), consiste em um conjunto de preditores em árvore de decisão, sendo que cada uma depende dos valores de um vetor aleatório amostrado independentemente e com a mesma distribuição para todas as árvores na floresta. O modelo RF usa um conjunto de árvores de decisão, onde cada nó interno representa um teste em uma característica (ou atributo) dos dados, cada ramificação representa o resultado do teste e cada nó folha representa uma classe ou valor de saída (BREIMAN, 2001; LIU *et al.*, 2022).

No estudo de Belgiu e Dragut (2016), foi destacado o sucesso do classificador RF no sensoriamento remoto, evidenciando sua capacidade de lidar com alta dimensionalidade e multicolinearidade dos dados. Adicionalmente, a investigação de Probst, Wright e Boulesteix (2019) sobre hiperparâmetros e estratégias de ajuste para *Random Forest* forneceu ideias sobre estratégias de otimização do desempenho do modelo na classificação de tipos de óleos lubrificantes.

Ainda sobre o modelo RF, a literatura é vasta em estudos demostrando sua aplicabilidade, como os estudos interdisciplinares, exemplificado pelo trabalho de Amor, Ghoul e Jemni (2023) intitulado "*Sign Language Recognition Using the Electromyographic Signal: A Systematic Literature Review*" e o artigo de Shao *et al.* (2023) intitulado "*The Application of Machine Learning Techniques in Geotechnical Engineering: A Review and Comparison*", para destacar a relevância do AM em contextos específicos do estudo. Da mesma forma, o estudo de Olu-Ajayi *et al.* (2023) intitulado "*Data-Driven Tools for Building Energy Consumption Prediction: A Review*", oferece ferramentas que orientam a aplicação eficaz de técnicas de AM, incluindo o modelo *Random Forest*, em previsões específicas.

Os parâmetros do modelo RF que são configurados e ajustados para personalizar o comportamento do algoritmo na *widget* "*Random Forest*" do Software Orange, são: (BREIMAN, 2001; DEMSAR *et al.*, 2013)

- **Número de árvores** (*Number of trees*): Este parâmetro determina quantas árvores de decisão serão incluídas na floresta. Um número maior

de árvores geralmente resulta em um modelo mais robusto, mas também pode aumentar o tempo de treinamento.

- **Número de árvores consideradas em cada divisão** (*Number of trees considered at each split*): Especifica quantos atributos serão aleatoriamente selecionados para consideração em cada nó ao fazer uma divisão. Isso introduz uma aleatoriedade adicional no processo de construção da árvore e pode ajudar a evitar o sobreajuste *(overfitting).*
- **Treinamento replicável** (*Replicable training*): Este parâmetro permite fixar a semente para a geração das árvores, garantindo a replicabilidade dos resultados. Permitindo a reprodução de experimentos ou resultados específicos.
- **Balanceamento da distribuição de classes** (*Balance class distribution*): Quando ativado, pondera as classes de forma inversamente proporcional às suas frequências. Sendo útil quando há classes desbalanceadas no conjunto de dados.

Controle de crescimento (*Growth control*):

- **Limitar a profundidade das árvores individuais** (*Limit depth of individual trees*): Permite especificar a profundidade máxima que as árvores individuais podem atingir. Ajudando a evitar o sobreajuste (overfitting), limitando a complexidade das árvores.
- **Não dividir subconjuntos menores que** (*Do not split subsets smaller than*): Define o tamanho mínimo do subconjunto que pode ser dividido. Controlando a granularidade das divisões nas árvores.

1.3.1.2 Máquina de Vetores de Suporte (SVM – *Support Vector Machines*)

O modelo Máquina de Vetores de Suporte (SVM – *Support Vector Machines*) trata-se de um algoritmo especial para classificação de dados e identificação de padrões. No estudo de Burges (1998) intitulado "*A Tutorial on Support Vector Machines for Pattern Recognition*", o SVM é apresentado como um modelo eficiente para trabalhar com problemas de classificação. Esta técnica de aprendizagem de Máquina opera construindo um hiperplano de decisão que otimiza a separação entre diferentes classes, tornando-a particularmente útil em cenários onde a complexidade das relações entre os dados é alta. O estudo mencionado aprofunda os conceitos teóricos necessários

para o funcionamento do SVM, destacando sua aplicabilidade em contextos de classificação binária (ADHIKARY *et al.*, 2022). A aplicabilidade do modelo SVM também é destacada em estudos como o de Vargas-Cardona *et al.* (2023), que examina a crescente interseção da IA em sua pesquisa sobre câncer, apresentando avanços na precisão diagnóstica de imagens para o rastreamento de câncer cervical, com ênfase no desempenho superior do SVM e métodos de AP (Aprendizagem Profunda).

Da mesma forma, Shao *et al.* (2023) revisaram a aplicação de algoritmos de AM, incluindo o SVM, na engenharia geotécnica, evidenciando a eficácia do SVM na classificação de tipos de solo e na previsão de propriedades geotécnicas. O estudo de Xu *et al.* (2023) contribuiu com uma abordagem do SVM para classificar tipos de óleo lubrificante com base em espectroscopia no infravermelho médio (MIR), obtendo resultados bastante precisos. Tais estudos demonstram a eficácia das técnicas de AM na análise de dados complexos e na previsão de desempenho em diversas áreas de aplicação.

Os seguintes parâmetros que são ajustáveis no algoritmo SVM são descritos a seguir (ADHIKARY *et al.*, 2022; BONESSO, 2013; DEMSAR *et al.*, 2013)

- **Custo** (Cost): Este parâmetro determina o termo de penalidade por erro do modelo SVM. Um valor maior de C significa que o modelo será menos tolerante a erros no conjunto de treinamento, o que pode levar a uma fronteira de decisão mais ajustada aos dados de treinamento, mas também pode aumentar o risco de sobreajuste (overfitting).
- **ε** (Epsilon): Este parâmetro é específico para regressão sendo utilizado no modelo ε-SVR. Ele define a distância dentro da qual nenhuma penalidade é aplicada aos valores previstos. Um valor maior de ε permite uma margem de erro maior nos resultados previstos.
- **v** (Nu): Este parâmetro é utilizado no modelo v-SVR e pode ser aplicado tanto em tarefas de classificação quanto de regressão. Ele define um limite superior na fração de erros de treinamento e um limite inferior na fração de vetores de suporte. Um valor menor de v indica uma maior restrição na quantidade de vetores de suporte permitidos.
- **Kernel**: O kernel é uma função que transforma o espaço de atributos em um novo espaço de características para se ajustar ao hiperplano de margem máxima. No widget SVM do Orange, podem ser escolhidos o kernel: Linear, Polinomial, RBF e Sigmóide. Cada kernel tem seus

próprios parâmetros, como a constante gama (**g**) para o kernel RBF e o grau do kernel (**d**) para o kernel polinomial.

1.3.1.3 Redes Neurais (NN - *Neural Network*).

As Redes Neurais (NN - *Neural Network*) são modelos computacionais inspirados na estrutura e funcionamento do cérebro humano, consistindo em camadas de unidades interconectadas, denominadas neurônios, capazes de aprender padrões complexos a partir de dados previamente fornecidos (HINTON; SALAKHUTDINOV, 2006). A utilização dessa técnica é respaldada por estudos relevantes na área do AM, como os trabalhos de Hinton e Salakhutdinov (2006) sobre a redução da dimensionalidade de dados com RN, sendo também abordada no trabalho de Kuppusamy, Nikolovski e Teekaraman (2023) na revisão de técnicas de AM para avaliação de desempenho de qualidade de energia em sistemas conectados à rede, e Huo *et al.* (2021) na previsão de desempenho de células de combustível de membrana de troca de prótons por meio de redes neurais convolucionais.

Os principais parâmetros ajustáveis em uma Rede Neural são (DEMSAR *et al.*, 2013; LIU *et al.*, 2018):

- **Neurônios por camada oculta** (*Neurons per hidden layer*): Este parâmetro permite definir o número de neurônios em cada camada oculta da RN. Cada elemento no parâmetro representa o número de neurônios em uma camada oculta específica. Escolher o número correto de neurônios por camada pode afetar significativamente o desempenho e a capacidade de generalização da RN.
- **Função de ativação para a camada oculta** (*Activation function for the hidden layer*): Esta opção permite selecionar a função de ativação (identidade, logística, tanh, ReLu) a ser usada nas camadas ocultas da RN. As funções de ativação determinam a não-linearidade da rede e sua capacidade de aprender padrões complexos nos dados.
- **Solver para otimização de peso** (*Solver for weight optimization*): Este parâmetro define o algoritmo de otimização usado para ajustar os pesos da RN durante o treinamento. Diferentes algoritmos de otimização têm diferentes características e podem ser mais adequados para diferentes tipos de problemas ou conjuntos de dados:

  - **L-BFGS-B**: um otimizador na família de métodos quase-Newton, sendo este método baseado em gradiente cujo objetivo é encontrar o valor ótimo de uma função, a fim de garantir a convergência do modelo (KUMAR et al., 2023).
  - **SGD**: descida do gradiente estocástico.
  - **Adam**: otimizador baseado em gradiente estocástico
- **Alpha**: Este parâmetro é o termo de penalização L2 (regularização) que ajuda a evitar o sobreajuste (*overfitting*), controlando a complexidade do modelo.
- **Máximo de iterações** (*Max iterations*): Define o número máximo de iterações (épocas) durante o treinamento da RN. Isso afeta quanto tempo o modelo é treinado e pode ajudar a evitar o *overfitting* se ajustado corretamente.

Essas referências proporcionam uma visão abrangente do estado atual e das perspectivas futuras do AM em diversas áreas, contextualizando-se dentro do escopo desta pesquisa.

A análise de óleos lubrificantes é uma prática essencial para garantir o desempenho e a durabilidade de motores automotivos (PASSONI; PACHECO; SILVEIRA, 2020; XU *et al.*, 2023). Através dessa análise, é possível identificar características importantes do óleo, como sua constituição base, aditivos e seu estado de degradação e contaminação por partículas indesejadas durante e após uso (BEZERRA *et al.*, 2021; GUAN *et al.*, 2011).

Diante do exposto, fica evidente que o AM aplicado na classificação de tipos de óleos lubrificantes de motores automotivos a partir de dados de espectroscopia Raman é uma abordagem promissora e com potencial para otimizar o processo de análise e identificação, podendo ser empregados nos dados espectrais como nos estudos de Passoni, Pacheco e Silveira (2020) e Bezerra *et al.* (2021). Com o avanço das tecnologias de coleta de espectros Raman e a crescente disponibilidade de dados espectrais e literatura da aplicação da espectroscopia Raman em análise de hidrocarbonetos, é possível explorar ainda mais as possibilidades dessa abordagem, contribuindo para aprimorar a eficiência e a confiabilidade da análise de óleos lubrificantes em benefício da indústria automotiva.

1.3.2. Métricas de avaliação em AM

A avaliação do desempenho dos modelos de AM na classificação de tipos de óleos lubrificantes de motores automotivos é essencial para determinar a confiabilidade e a eficácia desses modelos (PASSONI; PACHECO; SILVEIRA, 2020). Para isto, diversas métricas são empregadas para fornecer uma visão abrangente das capacidades de classificação dos modelos, entre elas, destacamos aqui a precisão (*Precision)*, a recall ou sensibilidade*,* o F1 *score,* a acurácia (*Accuracy*) e a especificidade (*Specificity*) (ZHOU *et al.*, 2023), fornecendo uma base sólida para avaliar comparativamente o desempenho dos modelos em questão*.*

A precisão representa a proporção de verdadeiros positivos em relação ao total de instâncias previstas como positivas, calculada pela fórmula descrita na Equação 2.1. Essa métrica é crucial quanto à minimização de falsos positivos e fundamental para a aplicação, como em situações em que erros de identificação positiva podem ter impactos significativos (WANG *et al.*, 2023).

$$Precisão = \frac{VP + FP}{VP} \tag{2.1}$$

O *recall*, também chamada de Sensibilidade, destaca a capacidade do modelo em identificar corretamente todas as instâncias positivas em relação ao total de instâncias positivas no conjunto de dados, calculada pela fórmula descrita na Equação 2.2. Em situações em que a identificação de todos os casos positivos é crucial, o *recall* torna-se uma métrica-chave a ser analisada (MARRS; NI-MEISTER, 2019).

$$Sensibilidade = \frac{VP}{VP + FN} \tag{2.2}$$

A F1 *score,* no que lhe concerne, busca um equilíbrio entre precisão e *recall*, calculada pela média harmônica entre essas duas métricas, conforme Equação 2.3. Essa métrica é particularmente útil em cenários de desequilíbrio de classe, proporcionando uma visão abrangente do desempenho do modelo (QURESHI *et al.*, 2023).

$$F1 = 2\,x\,\frac{Precisão\ x\ Sensibilidade}{Precisão + Sensibilidade} \quad (2.3)$$

Além destas métricas, a acurácia representa a proporção de instâncias corretamente classificadas em relação ao total de instâncias, calculada pela Equação 2.4. Fornecendo uma visão geral da precisão do modelo, indicando a proporção de predições corretas em relação ao total de predições (FENG; ZHENG; LIU, 2023):

$$Acurácia = \frac{VP + VN}{Total\ de\ Itens} \quad (2.4)$$

A especificidade destaca a habilidade do modelo em identificar corretamente as instâncias negativas (LEI *et al.*, 2020), calculada pela Equação 2.5:

$$Especificidade = \frac{VN}{VN + FP} \quad (2.5)$$

A análise dessas métricas proporciona uma compreensão abrangente do desempenho do modelo em diferentes aspectos da classificação, permitindo ajustes e otimizações para atender aos requisitos específicos da aplicação.

1.3.3 Matriz confusão multiclasse

Ao explorar padrões espectrais distintos associados a diferentes tipos de óleos lubrificantes, o AM pode ser empregado para a classificação multiclasse desses óleos em classes específicas, e o resultado da classificação pode ser interpretado através da matriz de confusão, compreendendo as classes de dados preditos comparativamente aos grupos reais como VP (Verdadeiro Positivo), VN (Verdadeiro Negativo), FP (Falso Positivo) e FN (Falso Negativo), conforme pode ser observado na Figura 1.

a)

Atual \ Predição	Mineral	Semissintético	Sintético
Mineral	VP	FN	FN
Semissintético	FP	VN	VN
Sintético	FP	VN	VN

b)

Atual \ Predição	Mineral	Semissintético	Sintético
Mineral	VN	FP	VN
Semissintético	FN	VP	FN
Sintético	VN	FP	VN

c)

Atual \ Predição	Mineral	Semissintético	Sintético
Mineral	VN	VN	FP
Semissintético	VN	VN	FP
Sintético	FN	FN	VP

Figura 1 - Exemplo de matrizes de confusão multiclasse para cada uma das 3 classes de lubrificantes: a) mineral, b) semissintético e c) sintético.

Em problemas multiclasse, não há uma única classe positiva e negativa; em vez disso, é necessário considerar cada classe individualmente. Para cada classe, os verdadeiros positivos (VP) são as instâncias corretamente previstas como pertencentes àquela classe, os verdadeiros negativos (VN) são as instâncias corretamente previstas como não pertencentes àquela classe, os falsos positivos (FP) são as instâncias incorretamente previstas como pertencentes àquela classe, e os falsos negativos (FN) são as instâncias que deveriam ter sido previstas como pertencentes àquela classe, mas foram incorretamente previstas como pertencentes a outras classes. Esse processo é realizado para cada classe individualmente, e os valores de VP, FP, VN e FN para cada classe são posteriormente juntados para calcular as métricas de desempenho global do modelo (PAGANO *et al.*, 2023).

Através das métricas definidas nas equações anteriores, neste trabalho serão comparados esses três modelos de AM para a classificação de tipos de óleos lubrificantes de motores automotivos. Visto que cada modelo possui suas vantagens e desvantagens, é importante considerar aspectos como desempenho, interpretabilidade dos resultados e requisitos computacionais ao escolher o modelo mais apropriado.

1.4 Justificativa

Conforme destacado por Passoni, Pacheco e Silveira (2020), a utilização do AM na classificação de tipos de óleos lubrificantes de motores automotivos a partir de dados de espectroscopia Raman apresenta-se como uma abordagem relevante. Uma vez que, a classificação precisa e eficiente desses óleos é essencial para garantir o desempenho adequado e prolongar a vida útil dos motores, além de contribuir para a redução da poluição ambiental e de danos ambientais causados por descartes inadequados (BEZERRA *et al.*, 2021; GUAN *et al.*, 2011). Portanto, novas pesquisas para efetivar investigações sobre procedimentos de AM em bancos de dados representados por sinais oriundos da espectroscopia Raman se mostram atualmente bastante necessárias.

E conforme demonstrado por Bezerra *et al.* (2021), a espectroscopia Raman é uma técnica altamente sensível e precisa para a análise molecular de substâncias, tendo se mostrado particularmente valiosa na avaliação das alterações químicas induzidas pela temperatura em óleos lubrificantes automotivos. Através da interação da radiação laser de excitação com a amostra a ser estudada, é possível obter um espectro característico que reflete a composição química e estrutural do material analisado (PASSONI; PACHECO; SILVEIRA, 2020).

O uso do AM permite explorar e extrair informações complexas contidas nos dados espectroscópicos de maneira eficiente como destacado em estudos anteriormente citados (PENG *et al.*, 2022; VARGAS-CARDONA *et al.*, 2023; YAN *et al.*, 2022). Assim como também, a aplicação do AM contribui para a melhoria dos processos de manutenção e inspeção automotiva. Uma vez que, a identificação correta do tipo de óleo lubrificante é fundamental para evitar danos aos motores, garantindo a segurança dos veículos (PASSONI; PACHECO; SILVEIRA, 2020). Com a utilização de algoritmos de classificação, é possível desenvolver sistemas inteligentes capazes de analisar instantaneamente os dados espectroscópicos de uma amostra de óleo, fornecendo uma resposta confiável sobre sua composição (GUAN *et al.*, 2011; XU *et al.*, 2023).

A utilização dessa abordagem acima citada com uso de AM permite explorar as informações contidas nos dados espectroscópicos, identificando padrões complexos e fornecendo uma classificação confiável quanto ao tipo ou qualidade dos óleos lubrificantes (PASSONI; PACHECO; SILVEIRA, 2020). Essa tecnologia pode contribuir para a melhoria dos processos de manutenção automotiva, garantindo a operação

adequada dos motores e reduzindo danos ambientais (BEZERRA *et al.*, 2021).

1.5 Objetivo

1.5.1 Objetivo geral

O objetivo geral desta dissertação é avaliar três tipos de modelos de AM, incluindo RF, SVM e RN, para a classificação de tipos de óleos lubrificantes de motores automotivos a partir de dados de espectroscopia Raman.

Portanto, essa dissertação visa explorar o uso do AM aplicado na classificação de tipos de óleos lubrificantes de motores automotivos a partir de dados de espectroscopia Raman. Serão investigados diferentes algoritmos e modelos de Aprendizado de Máquina Supervisionados, bem como técnicas de pré-processamento de dados, visando obter um modelo em modo supervisionado capaz de realizar uma classificação precisa e eficiente.

1.5.2 Objetivos específicos

Visando alcançar o objetivo geral do trabalho, teve-se como objetivos específicos:

a) Escolher e preparar um conjunto de dados representativos de espectroscopia Raman de óleos lubrificantes de motores automotivos, incluindo diferentes tipos e variações.

b) Implementar e ajustar modelos de RF, SVM e RN para a classificação dos diferentes tipos de óleos lubrificantes, a saber, em mineral, semissintético e sintético.

c) Avaliar o desempenho dos modelos usando métricas apropriadas, como acurácia, precisão, *recall* e F1-*score*.

d) Comparar os modelos, identificando seus pontos fortes e fracos como desempenho e interpretabilidade.

e) Fornecer uma discussão sobre os resultados obtidos, incluindo informações sobre a viabilidade do uso de técnicas de AM para a classificação de óleos lubrificantes de motores automotivos baseados em espectroscopia Raman.

2 MATERIAIS E MÉTODOS

2.1 Base de dados utilizada

Nesta pesquisa foram utilizados dados de espectroscopia Raman obtidos de análises em diferentes tipos de óleos lubrificantes para avaliar três tipos de modelos de Aprendizado de Máquina, RF, SVM e RN. Os dados de espectroscopia Raman foram cedidos pelo Laboratório de Espectroscopia Vibracional do Centro de Inovação, Tecnologia e Educação – CITE da Universidade Anhembi Morumbi, situado no Parque Tecnológico de São José dos Campos, SP.

Na coleta dos dados Raman foram usados três tipos de óleo lubrificante automotivos: mineral 15W30, semissintético 15W40 e sintético 5W30, da marca LUBRAX produzidos pela empresa Petrobras Distribuidora S.A, multiviscosos e recomendados para motores a gasolina, etanol, *flex* e gás natural veicular (GNV). Esse conjunto de dados faz parte de um experimento de identificação de degradação de óleos lubrificante quando aquecido por determinado período, a partir dos dados da dissertação de Bezerra (2020), intitulada “Análise da degradação de óleos lubrificantes em função da temperatura através da espectroscopia Raman”. No entanto, para o trabalho aqui apresentado, limitou-se apenas a classificação das amostras de óleos lubrificantes quanto ao tipo, a saber, mineral, semissintético e sintético, sem considerar o tempo de aquecimento.

Os espectros Raman foram obtidos em um espectrômetro Raman dispersivo (*Lambda Solutions,* Inc., MA, EUA, modelo P1) operando com um laser diodo de comprimento de onda de 830 nm e potência de 350 mW para excitação do espalhamento, com resolução espectral de 2 a 4 cm^{-1} na faixa espectral entre 400 e 1800 cm^{-1}. Os dados espectrais coletados passaram por um processo de pré-processamento, que incluiu remoção de ruídos de raios cósmicos (sinais aleatórios de alta intensidade que ocupam um ou dois pixels), remoção das linhas de base por ajuste e subtração de um polinômio de ordem 7 e normalização dos espectros pela área sob a curva.

Este conjunto de dados com 36 amostras para cada tipo (classe) de óleo lubrificante, totalizou 108 instâncias[1] ou observações individuais dos dados, medidos em diferentes intervalos de tempo na pesquisa de dissertação de Bezerra (2020).

Composto por 1050 atributos, sendo dessas 1049 características (*features*) numéricas, representando os valores do deslocamento Raman (número de onda) de cada espectro Raman característico de cada amostra, e 1 atributo nominal que é a variável alvo (*target*) dos tipos de óleos lubrificantes. Estas características dos dados Raman utilizados neste trabalho estão mostradas na Figura 2.

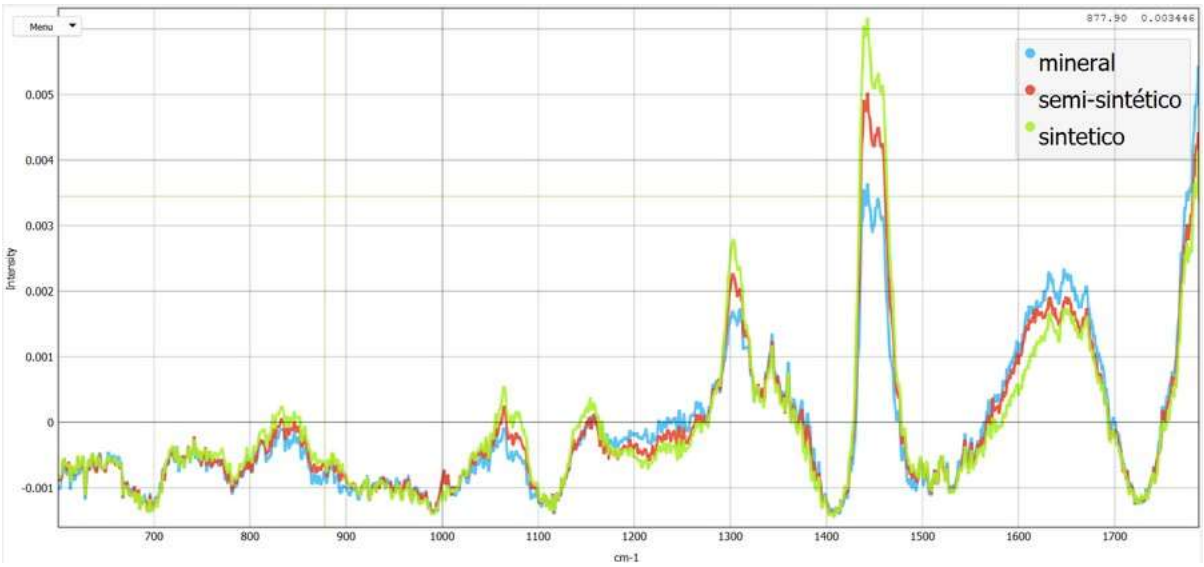

Figura 2 - Espectro Raman para 3 classes de óleo lubrificante automotivo.

Nesta pesquisa, os modelos de AM foram tratados como um problema de classificação multiclasse, onde são treinados para classificar os óleos lubrificantes com base em padrões complexos e sutis presentes nos espectros Raman dos diferentes grupos. A fim de desenvolver essa pesquisa, foi usado o ambiente Orange (Software Orange Data Mining, Bioinformatics Laboratory, Ljubljana, Eslovênia) (DEMSAR *et al.*, 2013).

[1] **instância** é um único exemplo de dados em um conjunto de treinamento ou teste, sendo cada linha do *dataset* das amostras nesta pesquisa e tem como característica possuir atributos que a descrevem (espectros) e um rótulo que indica sua classe (tipo de óleo).

2.2 Pré-processamento

Para a implementação dos modelos de AM foi aplicado um pré-processamento nos dados. Este procedimento foi necessário para preparar os dados brutos coletados por meio da espectroscopia Raman, e assim, ao combinar essas etapas de pré-processamento e análise, foi possível melhorar a precisão e a eficácia do AM na classificação de tipos de óleos lubrificantes de motores automotivos.

Inicialmente foi aplicada a normalização em relação à média, com o intuito de permitir a comparabilidade das características dos dados. E em seguida foi feita a implementação da análise de componente principal (PCA – *principal component analysis*). Esta etapa de pré-processamento com a PCA foi necessária para reduzir a dimensionalidade dos dados e identificar os principais componentes que explicam a variabilidade nos espectros Raman, conforme ilustrado na Figura 3.

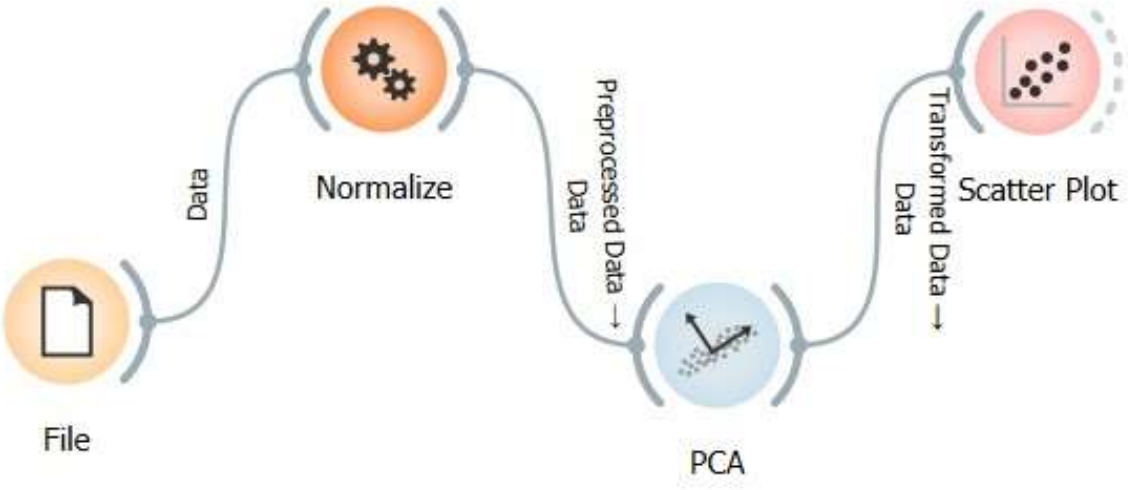

Figura 3 - Pré-processamento com Normalização e PCA.

Para aprimorar a consistência e comparabilidade dos dados, utilizou-se o *widget* "*Preprocess*" no ambiente Orange, selecionando a opção "*Normalize features – center to* μ". Essa etapa visava normalizar as características dos espectros em relação à média, promovendo uma homogeneização na escala vertical dos dados. Posteriormente, empregou-se o *widget* "*PCA* (*Principal Component Analysis*)" para realizar a redução de dimensionalidade, visando preservar informações significativas dos espectros enquanto se reduz a complexidade do conjunto de dados, preparando-o de para a aplicação de modelos do AM, na tarefa de classificação.

2.3 Parametrização dos modelos de classificação no software Orange

Conforme descrito no capítulo anterior, neste trabalho foram examinados três modelos de AM supervisionados para a classificação desses óleos lubrificantes: Floresta Aleatória (RF *Random Forest*), Máquina de Vetores de Suporte (SVM - *Support Vector Machine* – SVM) e Redes Neurais (NN - *Neural Network*). Estes modelos foram implementados e simulados no ambiente Orange.

2.3.1. Parametrização do modelo RF

Conforme foi descrito na Fundamentação Teórica, o modelo RF é um algoritmo de aprendizado supervisionado que utiliza uma combinação de árvores de decisão para realizar a classificação, conseguindo lidar com dados multidimensionais e não lineares, o que o torna uma escolha promissora para a classificação dos tipos de óleos lubrificantes.

No modelo RF apresentado, foram descritos os parâmetros que são configurados e ajustados para personalizar o comportamento do algoritmo na *widget* "*Random Forest*". Neste estudo onde são utilizados dados de espectroscopia Raman de diferentes tipos de óleos lubrificantes foram configurados no software Orange os três primeiros parâmetros desta *widget* "*Random Forest*". A saber: Número de árvores, Número de árvores consideradas em cada divisão e Treinamento replicável.

2.3.2. Parametrização do modelo SVM

O segundo modelo utilizado na simulação de AM foi o SVM que, conforme foi descrito na Fundamentação Teórica, trata-se de um algoritmo de aprendizado supervisionado que mapeia os dados em um espaço de alta dimensionalidade e encontra um hiperplano que separa as diferentes classes. Assim o objetivo é encontrar o melhor limite de decisão, maximizando a margem entre as classes. Conforme supracitado foram mostrados os principais parâmetros de ajuste do algoritmo SVM.

Os parâmetros no algoritmo SVM na *widget* "*SVM*" do software Orange configurados neste estudo, onde são utilizados dados de espectroscopia Raman de diferentes tipos de óleos lubrificantes, foram: Custo, *kennel*, a Tolerância numérica e o Limite de iterações.

2.3.3. Parametrização do modelo RN

O terceiro modelo implementado foi o RN, que tem se mostrado muito eficaz na classificação de dados complexos. Neste trabalho foi utilizado o *widget* "*Neural Network*" do software Orange que constrói e treina redes neurais artificiais utilizando o algoritmo de *perceptron* de múltiplas camadas (MLP) com retropropagação (*backpropagation*). Dessa forma é permitido a criação de modelos não-lineares capazes de aprender relações complexas nos dados, no entanto, é importante ajustar adequadamente a arquitetura da RN e otimizar seus parâmetros para obter os melhores resultados. Os principais parâmetros de uma RN foram descritos na Fundamentação Teórica.

Neste estudo, onde são utilizados dados de espectroscopia Raman de diferentes tipos de óleos lubrificantes, foram configurados os seguintes parâmetros da *widget* "*Neural Network*" do software Orange: Neurônios por camada oculta, Função de ativação para a camada oculta, *Solver* para otimização de peso, Alpha e o Máximo de iterações.

2.4 Treinamento e teste

A fim de dividir os dados em conjuntos de treinamento e teste, permitindo que se avaliem a capacidade de generalização dos modelos de classificação escolhidos, fora usado a *widget* "*Test and score*". Além disso, nesta pesquisa foi utilizada a validação cruzada, que envolve dividir o conjunto de dados em múltiplas partições, treinar o modelo em diferentes combinações de partições e calcular a média dos resultados para obter uma estimativa mais robusta do desempenho do modelo.

2.5 Métricas de avaliação do desempenho

As métricas usadas em classificação, *widget "Test and score*" para avaliar o desempenho do modelo foram: precisão, sensibilidade, especificidade e F1-*score*. Considerou-se na seleção que essas métricas fornecem uma compreensão abrangente do quão bem o modelo está performando em sua tarefa de classificação.

3 RESULTADOS E DISCUSSÃO

O esquema de teste dos modelos implementados nesta dissertação para avaliar o desempenho dos classificadores de AM são vistos na Figura 4. Conforme definido no capítulo anterior, inicialmente, os dados brutos de espectroscopia Raman foram submetidos a um processo de normalização e, em seguida, ao método da PCA para redução de dimensionalidade.

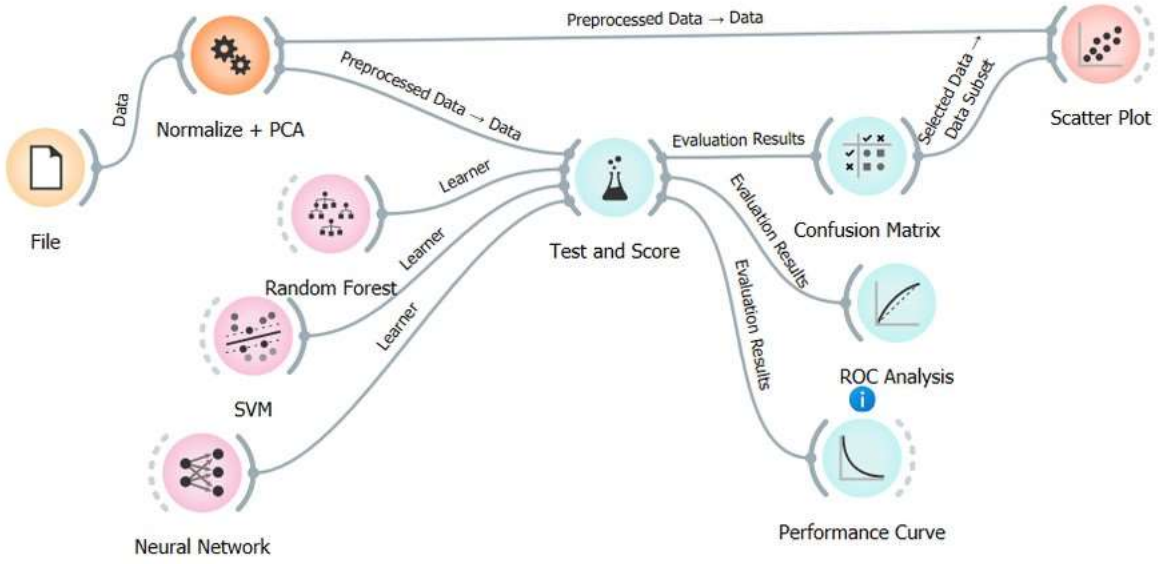

Figura 4 - Esquema de teste dos modelos e visualização dos resultados.

Esse processo de pré-processamento teve como objetivo extrair as principais características dos dados, proporcionando uma representação mais eficiente para a implementação dos modelos de AM. O fluxo inicial dos dados é mostrado na Figura 4 onde se verifica que os modelos *Random Forest*, SVM e *Neural Network* foram integrados ao *widget* "*Test and score*" para a avaliação do desempenho dos modelos. A escolha desses modelos proporcionou uma variedade de abordagens de AM, permitindo uma comparação abrangente de suas capacidades em classificar diferentes tipos de óleos lubrificantes. A utilização de validação cruzada com estratificação de 10 *folds* garantiu uma avaliação robusta e representativa, considerando a variabilidade presente nos dados.

Após a fase de teste do modelo, os resultados foram encaminhados para o *widget* "*Confusion Matrix"*, que proporcionou uma visão detalhada da matriz de confusão para cada modelo multiclasse, indicando as taxas de VP, VN, FP e FN. Para

uma compreensão visual mais aprofundada, os resultados foram plotados em um gráfico de dispersão. Neste gráfico estão destacados particularmente os resultados falsos, permitindo uma análise visual dos casos em que o modelo não conseguiu classificar corretamente os óleos lubrificantes. Essa abordagem combinada de pré-processamento, testes, visualização e análise dos resultados contribui significativamente para a avaliação abrangente e compreensão do desempenho dos modelos de AM aplicados a dados de espectroscopia Raman na classificação de tipos de óleos lubrificantes automotivos.

3.1 Pré-processamento dos dados

Como destacado no capítulo anterior, a normalização foi empregada a fim de aprimorar a consistência e comparabilidade dos dados, utilizou-se o *widget* "*Preprocess*" no ambiente Orange, selecionando a opção "*Normalize features – center to* μ". Essa etapa visou normalizar as características dos espectros em relação à média, promovendo uma homogeneização na escala dos dados. Posteriormente, empregou-se o *widget* "*PCA* (*Principal Component Analysis*)" para realizar a redução de dimensionalidade. Foram escolhidos 3 PCs, que asseguraram uma explicação de variância acumulada de 83%, conforme a Figura 5.

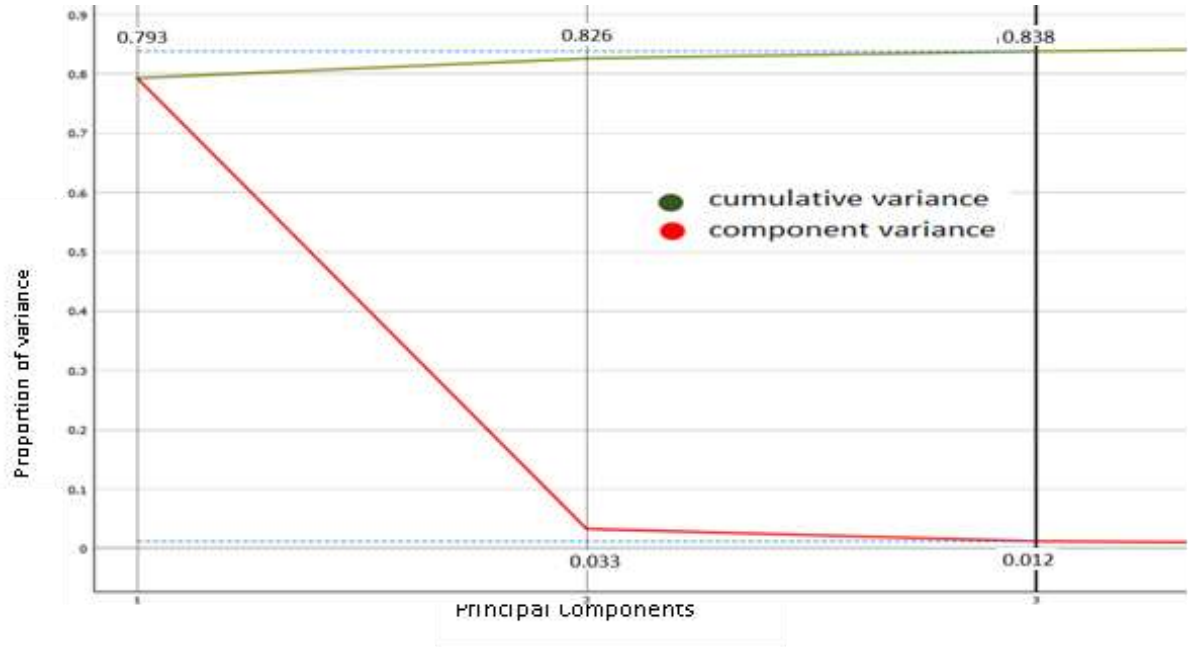

Figura 5 - Porcentagem de variância explicada por cada componente principal.

Esta escolha de PCs visou preservar informações significativas dos espectros enquanto reduzia a complexidade do conjunto de dados, preparando-o de maneira eficaz para a aplicação de modelos de AM na tarefa de classificação.

O gráfico de dispersão e em particular, a projeção no plano definido pelos Componentes Principais 2 e 3 (PC2) e PC3) revelou uma clara e nítida diferenciação de classes para cada tipo de óleo lubrificante, conforme mostrado na Figura 6. Veja apêndice B, as projeções (PC1 x PC2) e (PC1 x PC3).

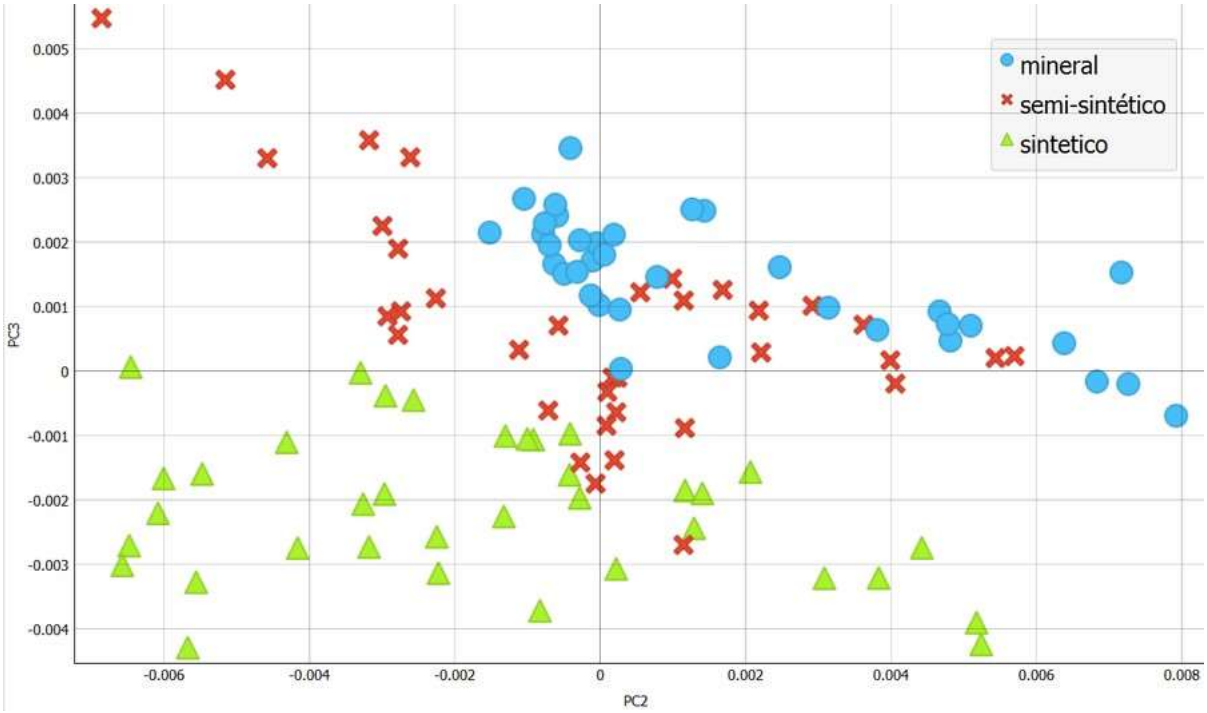

Figura 6 - Projeção de Componentes Principais 2 e 3 (PC2 e PC3) com melhor separação das classes.

3.2 Configuração dos modelos de AM

Neste estudo foram configurados os três primeiros parâmetros da *widget* "*Random Forest*": O **Número de árvores**, que teve melhor desempenho foram 40 árvores, conforme as curvas de validação nos Apêndices C.1 e C.2. Uma vez que, um número maior de árvores geralmente leva a um modelo mais robusto, embora possa aumentar o tempo de treinamento (PROBST; WRIGHT; BOULESTEIX, 2019). O **Número de árvores consideradas em cada divisão** foram 3 correspondendo as 3 classes de óleo lubrificante. Fora também selecionado o parâmetro, **Treinamento replicável**, a fim de fixar a semente para a geração das árvores, garantindo a

replicabilidade dos resultados, veja o Apêndice A.1.

Para o modelo SVM, visando otimizar o desempenho do modelo, seguindo as recomendações de ajuste de hiperparâmetros propostas por Adhikary *et al.* (2022). Foram configurados os parâmetros da *widget* "*SVM*": **custo (Cost)** foi fixado em 0,84, a fim de permitir uma ponderação equilibrada entre a precisão do modelo e a generalização para novos dados, conforme Apêndice C.4. A **margem de erro (ε)** foi estabelecida em 0,10, indicando uma tolerância aceitável para variações nos dados. O ***Kernel*** escolhido foi o "*Linea*r", sugerindo uma abordagem linear na separação das classes, veja o Apêndice C.3. Além disso, foram definidos uma **tolerância numérica** de 0,001 para garantir a estabilidade computacional e um **limite de iteração** de 100 para evitar possíveis problemas de convergência. Essa parametrização buscou assegurar maximização da eficácia e a precisão do modelo, conforme Apêndice A.2.

Na parametrização do modelo da RN, conforme curvas de validação no Apêndice C.5, usando a *widget "Neural Network",* foram escolhidos 1 neurônios na **camada oculta**, buscando evitar tanto a simplicidade que pode levar a subajuste (*underfitting*) quanto a complexidade excessiva que pode induzir ao sobreajuste (*overfitting*) (MARRS; NI-MEISTER, 2019). A utilização da **função de ativação** "*Identity*" para as camadas ocultas permitiu a preservação dos dados, potencialmente capturando padrões mais sutis e refinados presentes na espectroscopia Raman após o pré-processamento com a PCA. No modelo RN, veja o Apêndice C.5, o **otimizador de peso** escolhido foi o L-BFGS-B, enquanto um **valor de "*alpha*"** estabelecido em 0,1 contribui para a regularização do aprendizado, evitando possíveis problemas de sobreajuste (*overfitting*). Além disso, fixou-se o **número máximo de interações** em 200 para atender à necessidade de compromisso entre eficiência computacional e a busca por convergência adequada durante o treinamento da RN, visando alcançar um equilíbrio entre a capacidade preditiva e a generalização do modelo, veja o Apêndice A.3.

Na próxima seção são apresentados os resultados obtidos pelos algoritmos classificadores (*Random Forest*, SVM, *Neural Network*) para classificar os dados de espectroscopia Raman nas classes: mineral, semissintético e sintético.

3.3 Implementação do algoritmo RF

Na Figura 7a é apresentada a matriz de confusão multiclasse, resultante da

aplicação do modelo RF na classificação de óleos lubrificantes, onde se revelaram padrões distintos de acertos e erros, e na Figura 7b é apresentada a contagem das classes preditas VP, VN, FP e FN.

a)

		Predição			
	Random Forest	Mineral	Semissintético	Sintético	
Atual	Mineral	30	6	0	36
	Semissintético	3	31	2	36
	Sintético	0	3	33	36
		33	40	35	108

b)

Random Forest	Mineral	Semissintético	Sintético
VP	30	31	33
VN	69	63	70
FP	3	9	2
FN	6	5	3
(VP + VN)	99	94	103
(FP + FN)	9	14	5

Figura 7 - a) Matriz de confusão do modelo Random Forest; b) Detalhamento da matriz de confusão.

Os resultados da matriz de confusão[2] para o modelo de classificação RF revelam informações sobre a capacidade do modelo em distinguir entre os diferentes tipos de óleos, conforme Figura 7a.

Considerando a classificação para óleo mineral observa-se que das 36 instâncias[3] o modelo RF classifica corretamente 30 como VP e classifica corretamente 69 (VN) instâncias como não pertencente a classe de óleo mineral. No entanto, em 3 (FP) instâncias o modelo classifica erroneamente como mineral o óleo semissintético e em 6 (FN) instâncias erra ao classificar como óleo semissintético instâncias que eram óleos minerais.

Analisando a classe de óleos semissintético, o modelo RF classificou corretamente 31 como VP e 63 com VN, consoante a Figura 7a. Mas errou ao classificar como semissintético em 9 (FP) instâncias de outro tipo de óleo, a saber, 6 (FP) mineral e 3 (FP) sintético. E em 5 (FN) instâncias o modelo errou em classificar como não semissintéticos óleos que eram semissintéticos, como se vê na Figura 7b.

---

[2] Veja a seção 2.6 acima: **Matriz confusão multiclasse**

[3] Veja a explicação de **instâncias** na nota 1.

Verificando a classe de óleos sintéticos no modelo RF, Figura 7b, observa-se que das 36 instâncias o modelo RF classifica corretamente 33 como VP como pertencente a classe dos sintéticos e em 70 (VN) instâncias que o modelo acerta como não pertencente a classe de óleo sintéticos. E em 2 (FP) instâncias o modelo classifica erroneamente como sintético óleo semissintético e em 3 (FN) instâncias erra em classificar como não óleo sintético instâncias de óleo sintético.

Ao considerar eficácia global do modelo RF para as 3 classes de tipos de óleos lubrificantes, o óleo sintético foi o melhor classificado, pois foi classificado corretamente em 103 (VP + VN) instâncias das 108 instâncias, seguido pelo óleo mineral 99 (VP + VN) e 94 (VP + VN) no óleo semissintético, Figura 7b.

O resultado da matriz de confusão é melhor destacado como visto na Figura 8 pela plotagem da dispersão entre as variáveis PC2 X PC3, com destaque nos FNs do modelo RF, sendo 6 (mineral), 5 (semissintético) e 3 (sintético), conforme Figura 7b.

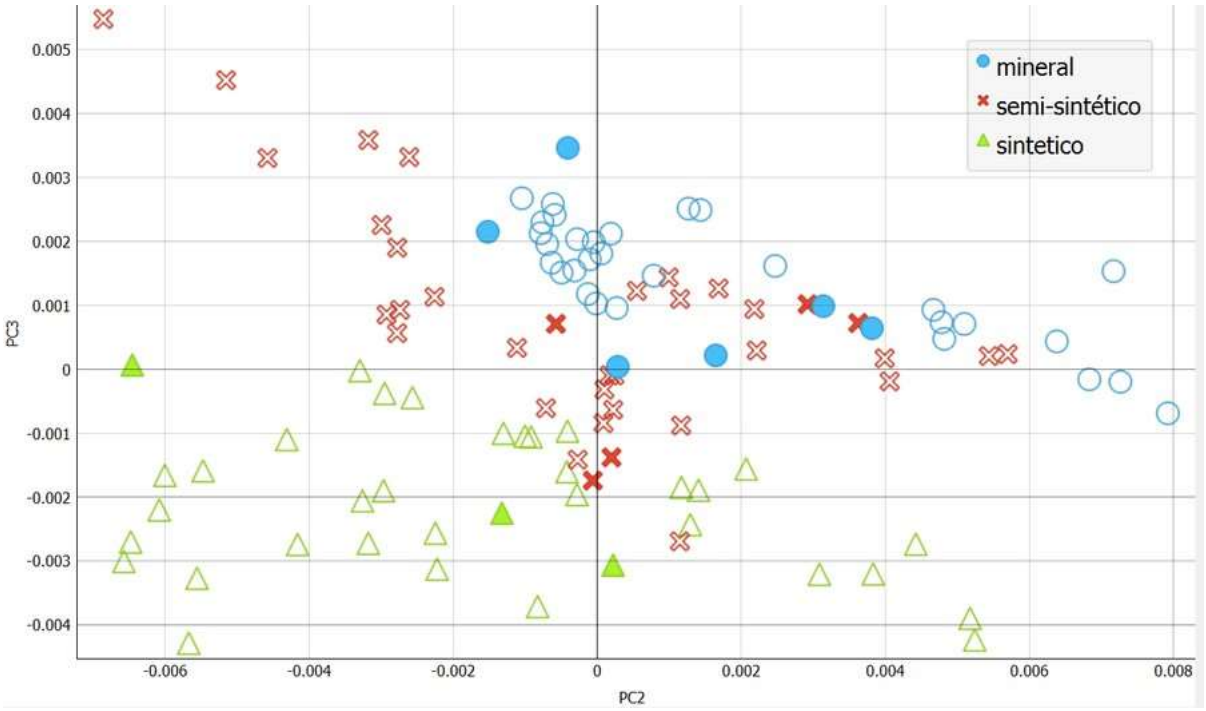

Figura 8 - Plotagem binária dos valores das variáveis PC2 X PC3 com destaque no FN do modelo Random Forest.

A análise da matriz de confusão destacou a eficácia do modelo RF na classificação de tipos de óleos lubrificantes, evidenciando sua capacidade em classificar corretamente a maioria dos casos. Contudo, a atenção às ocorrências de FP e FN é crucial para aprimorar a precisão do modelo, contribuindo para a confiabilidade

nas aplicações práticas de AM na indústria de lubrificantes automotivos.

3.4 Implementação do algoritmo SVM

A análise da matriz de confusão para o modelo o SVM na classificação de tipos de óleos lubrificantes ofereceu uma visão detalhada do desempenho do modelo em diferentes classes dos óleos lubrificantes. Observou-se que o SVM alcançou um número significativo de verdadeiros positivos (VP) em todas as classes, destacando sua habilidade em identificar corretamente os tipos de óleos em questão. Notavelmente, a classe sintético apresentou o maior número de VP, 36 VP de 36 instâncias, indicando uma maior acurácia na identificação dessa classe específica de óleo, seguido pelo óleo mineral 34 (VP) e o óleo semissintético 31 (VP), conforme Figura 9a.

a)

		Predição			
	SVM	Mineral	Semissintético	Sintético	
Atual	Mineral	34	2	0	36
	Semissintético	5	31	0	36
	Sintético	0	0	36	36
		39	33	36	108

b)

SVM	Mineral	Semissintético	Sintético
VP	34	31	36
VN	67	70	72
FP	5	2	0
FN	2	5	0
(VP + VN)	101	101	108
(FP + FN)	7	7	0

Figura 9 - a) Matriz de Confusão do modelo SVM; b) Detalhamento da matriz de confusão.

Na análise da taxa de verdadeiros negativos (VN) na Figura 9b, observou-se que para os óleos sintéticos o modelo SVM classificou corretamente 72 (VN) instâncias como não pertencente a essa classe de óleo e modelo SVM também mostrou um desempenho consistente, nas classes mineral e semissintético, com valores de VN 67 e 70 respectivamente. Isso sugere que o modelo consegue distinguir eficientemente entre as classes específicas de óleos.

A robustez geral do modelo na correta classificação dos óleos lubrificantes é observada na soma dos (VP + VN) para cada classe de óleo, com diferencial para o óleo sintético onde acertou todas as 108 instâncias, seguido pelos óleos minerais e

semissintético onde ambos acertaram corretamente 101 instância. No entanto, para óleos minerais o modelo apresentou 5 (FP) e 2 (FN), e para os semissintéticos, 2 (FP) e 5 (FN). A soma desses valores (FP + FN) proporciona uma visão abrangente dos erros cometidos na classificação desses óleos, sendo 7 em ambos os óleos, veja a Figura 9b. No diagrama de dispersão, Figura 10, entre as componentes principais PC2 x PC3, destaca-se os FNs na classificação dos óleos mineral e sintético.

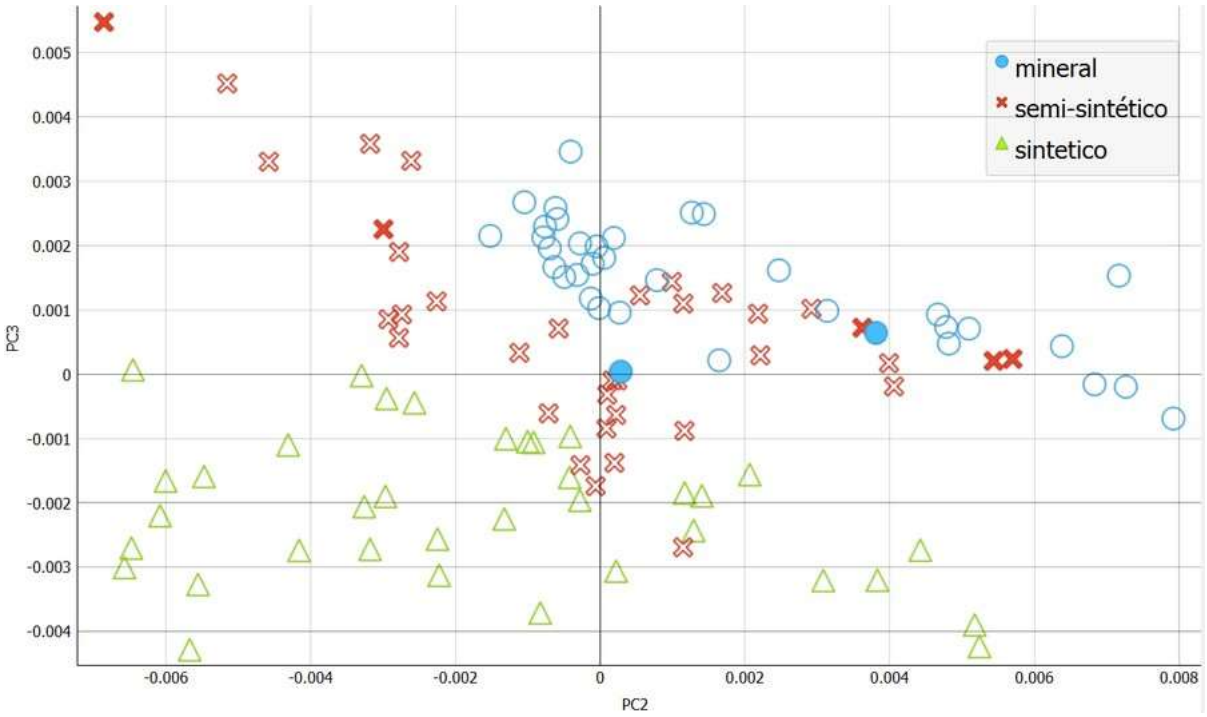

Figura 10 - PC2 x PC3 com destaque no False Negative do modelo SVM.

Em suma, a análise detalhada da matriz de confusão do modelo SVM destaca o alto desempenho na classificação de óleos lubrificantes, proporcionando informações valiosas de sua utilização em associação com a espectroscopia Raman.

3.5 Implementação do algoritmo RN

Os resultados da matriz de confusão para o modelo de classificação RN revelam a eficiência do modelo em distinguir entre os diferentes tipos de óleos, conforme Figura 11.

Observamos que o óleo sintético classificou corretamente todas as 36 (VP) instâncias, enquanto os óleos minerais e semissintéticos ambos foram 32 (VP) de cada

uma das 36 instâncias. Na análise da taxa de verdadeiros negativos (VN) na Figura 11b, observou-se que para os óleos sintéticos o modelo RN classificou corretamente 72 (VN) instâncias como não pertencente a essa classe de óleo e modelo RN também mostrou um desempenho consistente, nas classes mineral e semissintético, ambos com 68 (VN). Isso sugere que o modelo consegue distinguir eficientemente entre as classes específicas de óleos.

Considerando os valores onde o modelo classificou corretamente a classe (VP) e os valores corretamente classificados como não pertencente as classes (VN), temos a soma (VP + VN) que dá uma visão da eficácia global do modelo: sintético 108 instâncias corretas, seguidos pelos óleos mineral e semissintético ambos 100 das 108 instâncias classificadas corretamente, como se vê na Figura 11b. Novamente a classe sintética no modelo RN, assim como no modelo SVM, teve ausência de falsos negativos (FN) e falsos positivos (FP), indicando uma precisão notável e destacando a robustez do modelo em evitar classificações erradas nessa classe de óleo.

a)

	Neural Network	Predição: Mineral	Predição: Semi-sintético	Predição: Sintético	
Atual	Mineral	32	4	0	36
	Semi-sintético	4	32	0	36
	Sintético	0	0	36	36
		36	36	36	108

b)

Neural Network	Mineral	Semissintético	Sintético
VP	32	32	36
VN	68	68	72
FP	4	4	0
FN	4	4	0
(VP + VN)	100	100	108
(FP + FN)	8	8	0

Figura 11 - a) Matriz de Confusão do modelo Neural Network; b) Detalhamento da matriz de confusão.

Conforme a Figura 11b, notou-se que, para óleos minerais e semissintéticos, o modelo apresentou 4 FP e 4 FN em ambos os casos. A soma desses valores (FP + FN) proporciona uma visão abrangente dos erros de classificação, sendo 8 em ambas as classes. No diagrama de dispersão, Figura 12, entre as componentes principais PC2 X PC3, destaca-se os FNs no modelo RN.

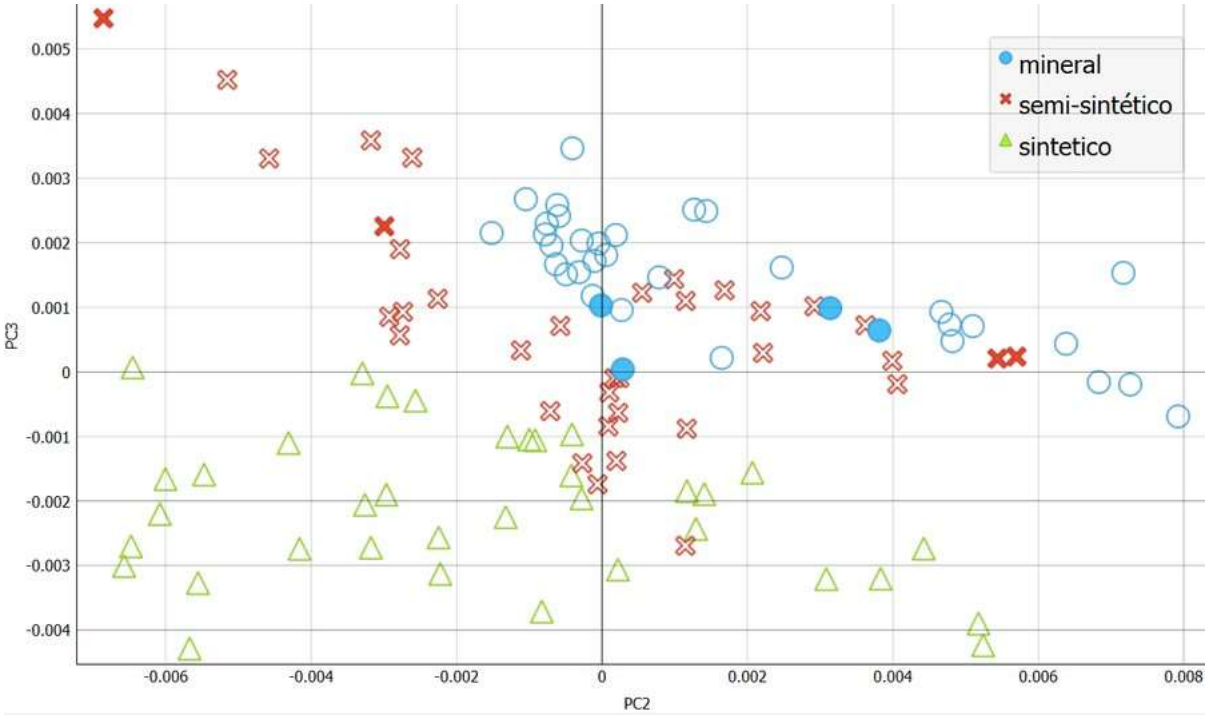

Figura 12 - PC2 x PC3 com destaque no False Negative do modelo Neural Network.

O desempenho no modelo RN sugere que a abordagem de AM, utilizando de redes neurais, é confiável para a classificação de tipos de óleos lubrificantes com base em dados de espectroscopia Raman.

3.6 Comparação entre os modelos dado o tipo de óleo:

A aplicação de AM na classificação de tipos de óleos lubrificantes de motores automotivos por meio de espectros Raman é um campo de pesquisa promissor para otimizar a identificação e diferenciação de diferentes composições de óleos lubrificantes. Nesse contexto, a partir da matriz de confusão explicada anteriormente nas Figuras 7, 9 e 11, avaliou-se o desempenho dos modelos de classificação (RF, SVM e RN) para cada tipo de óleo lubrificante, com as métricas a seguir: precisão (*Precision*), recall ou Sensibilidade, F1 *score*, acurácia (*Accuracy*) e especificidade (*Specificity*), veja seção 2.5.

3.6.1 Óleo mineral

A análise das métricas de desempenho para a classificação do óleo lubrificante mineral revela nuances distintas entre os modelos de classificação RF, SVM e RN. No

que diz respeito à métrica de precisão, que mensura a proporção de instâncias classificadas como positivas corretamente em relação ao total de instâncias classificadas como positivas, observou-se que o RF alcançou 90,9%, indicando uma alta precisão na identificação de óleos minerais. A RN e o SVM apresentaram valores ligeiramente menores, com 88,9% e 87,2%, respectivamente, conforme a Figura 13.

Mineral	Random Forest	SVM	Neural Network
Precision	90,9%	87,2%	88,9%
Recall	83,3%	94,4%	88,9%
F1	87,0%	90,7%	88,9%
Accuracy	91,7%	93,5%	92,6%
Specificity	95,8%	93,1%	94,4%

Figura 13 - Comparação entre os modelos de classificação para o óleo mineral.

O *recall*, que avalia a proporção de instâncias positivas corretamente identificadas em relação ao total de instâncias positivas, conforme Figura 13, demonstrou que o SVM obteve o melhor desempenho, atingindo 94,4%, indicando uma capacidade superior em classificar óleos minerais. A RN e o SVM, embora apresentando bons resultados, registraram valores de 88,9% e 83,3%, respectivamente. E conforme a Figura 14, ao avaliar o gráfico da precisão *versus recall*, é possível observar que todos os modelos de classificação apresentam resultados superiores a 80%.

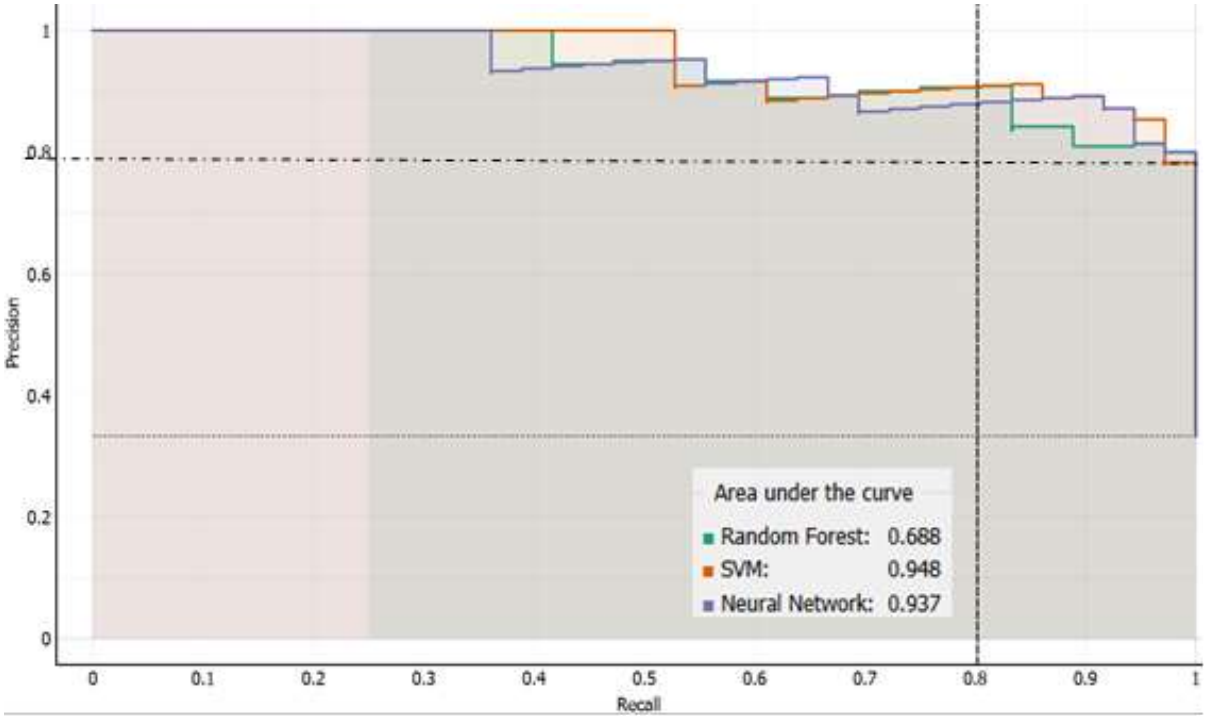

Figura 14 - Precisão versus recall, maior que 80% para todos os modelos na classificação de óleos minerais.

A métrica F1 *score*, que combina precisão e *recall*, revela a harmonia entre essas duas medidas. O SVM, mais uma vez, se destaca com um valor de 90,7%, seguido pelo RN com 88,9%, e o RF com 87,0% como de observa na Figura 13. A acurácia, que mede a proporção de predições corretas em relação ao total de instâncias. A Figura 13 indicou que todos os modelos apresentaram resultados favoráveis, com destaque para o SVM atingindo 93,5%. A RN e o RF alcançaram valores de 92,6% e 91,7%, respectivamente.

Por fim, a especificidade, que avalia a proporção de instâncias negativas corretamente identificadas em relação ao total de instâncias negativas, mostrou que todos os modelos tiveram desempenho satisfatório, com destaque para o RF registrando 95,8%, seguido de 94,4% para a RN e 93,1% no SVM , indicando sua habilidade em distinguir óleos não minerais, como se pode observar na Figura 13.

Ao realizar uma análise comparativa das métricas de classificação no modelo RF para óleos lubrificantes minerais, nota-se conforme a Figura 15 que todos eles apresentaram um bom desempenho. Entretanto, é evidente que o modelo SVM destacou-se como o mais eficaz, liderando em três das cinco métricas analisadas, nomeadamente *recall* (94,4%), acurácia (93,5%) e F1 *score* (90,7%). Seguido do modelo RF que demonstrou superioridade em duas das cinco métricas, a saber, na

especificidade (95,8%) e na precisão (90,9%).

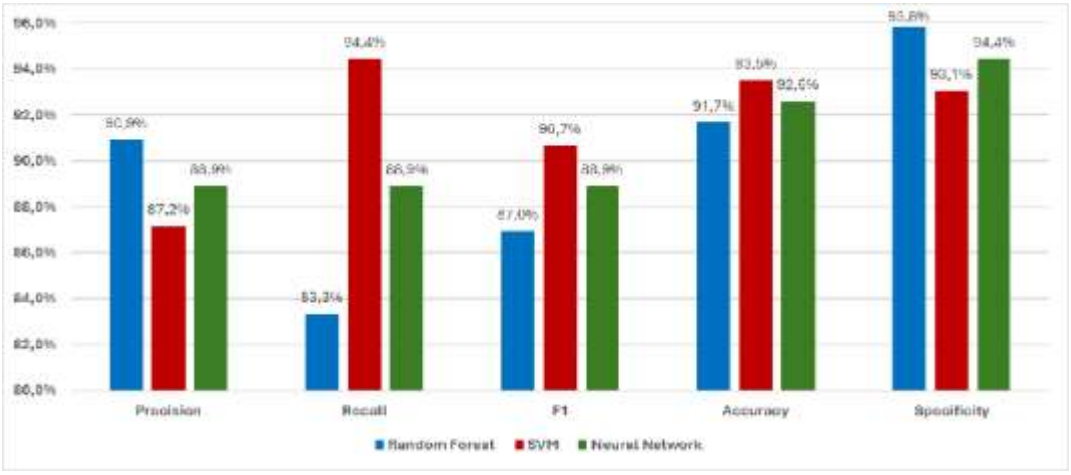

Figura 15 - Comparação entre as métricas de classificação para o óleo mineral.

Esta análise revela que cada modelo exibe pontos fortes em métricas específicas, proporcionando uma visão mais abrangente e informada sobre o desempenho global de cada algoritmo na tarefa de classificação de óleos lubrificantes minerais.

3.6.2 Óleo semissintético

No caso dos óleos semissintético, ao consideramos a métrica precisão na Figura 16 que indica a proporção de instâncias positivas corretamente identificadas em relação ao total de instâncias classificadas como positivas, o modelo SVM e RN tiveram os resultados notavelmente mais altos, atingindo 93,9% e 88,9%, respectivamente, demonstrando uma capacidade superior de evitar falsos positivos. Enquanto o RF alcançou 77,5% das classificações positivas como corretas.

Semis-sintético	Random Forest	SVM	Neural Network
Precision	77,5%	93,9%	88,9%
Recall	86,1%	86,1%	88,9%
F1	81,6%	89,9%	88,9%
Accuracy	87,0%	93,5%	92,6%
Specificity	87,5%	97,2%	94,4%

Figura 16 - Comparação entre os modelos de classificação para o óleo semissintético.

Analisando a métrica *recall*, que destaca a capacidade do modelo em identificar todas as instâncias positivas. Observa-se que a RN teve o maior *recall* alcançando 88,9%. Indicando uma robustez consistente na identificação de instâncias positivas para os óleos semissintéticos. Enquanto os modelos RF e o SVM apresentam um *recall* de 86,1%, como bem detalhado na Figura 16.

Considerando na Figura 16, o F1-*score*, sendo uma média harmônica de precisão e *recall*, que fornece uma visão equilibrada do desempenho do modelo. Observasse que o óleo semissintético obteve 89,9% tanto no modelo SVM seguido pelo Modelo RN com 88,9%, indicando um equilíbrio satisfatório entre precisão e *recall* para esses algoritmos e 81,6% para o RF. E considerando o gráfico na Figura 17 entre a precisão *versus recall*, pode se ver os modelos SVM e *Neural Network*, ambos são maiores que 80%, ressaltando o equilíbrio entre precisão e *recall*.

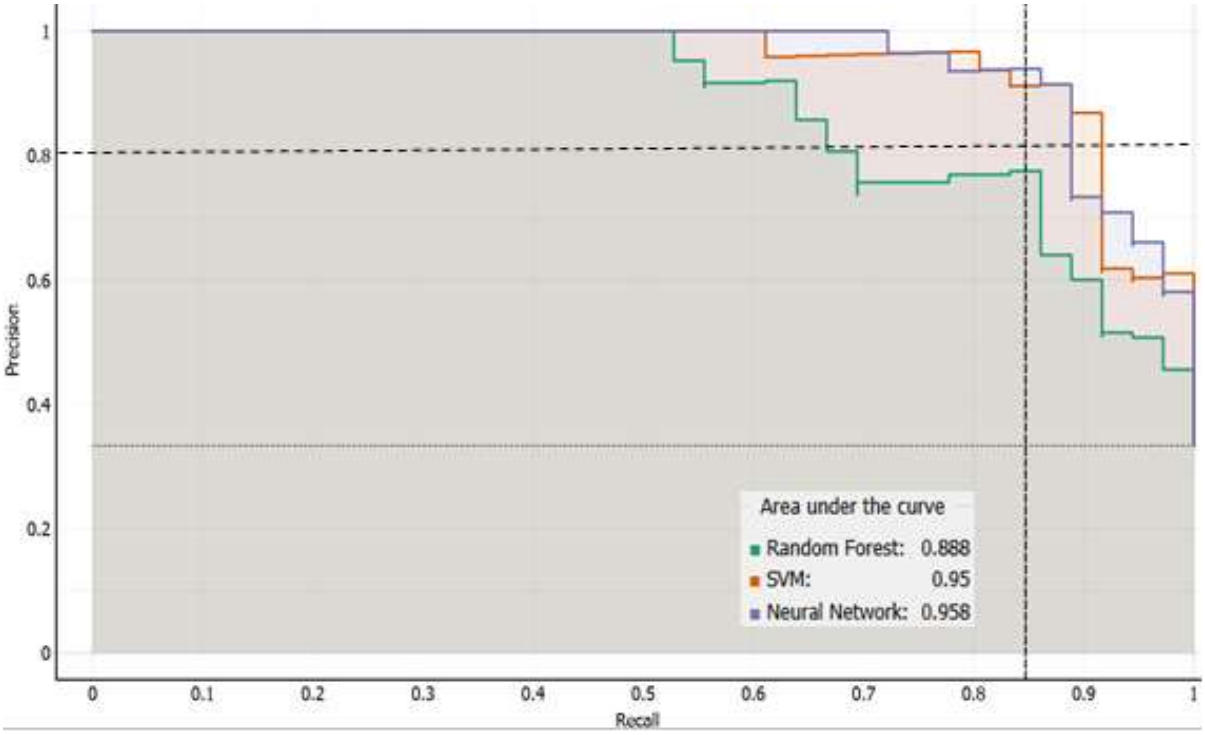

Figura 17 - Precisão versus recall, maior que 80% para os modelos SVM e RN na classificação de óleos semissintéticos.

A próxima métrica a acurácia representa a proporção total de previsões corretas feitas pelo modelo, independente da classe. Os resultados da Figura 16, indicam um desempenho geral em média de 87,0% para o óleo semissintético, sendo que o modelo SVM lidera com uma acurácia de 93,5%, demonstrando sua eficácia geral na classificação dos óleos lubrificantes com alta capacidade de fazer previsões corretas. Logo em seguida temos com 92,6% a RN e 87,0% para o RF.

Por fim, avaliando a capacidade do modelo em identificar corretamente as instâncias negativas através da métrica especificidade. O modelo SVM destaca-se com uma especificidade de 97,2% superior aos demais, indicando uma habilidade excepcional em evitar falsos positivos e classificar corretamente os óleos não pertencentes à classe em questão. No entanto, os outros modelos tiveram bom desempenho a RN (94,4%) e o RF (87,5%), veja a Figura 16.

Em resumo, a análise das métricas destacam a robustez do modelo SVM na classificação de óleos semissintético em diversos aspectos em relação aos outros modelos, liderando em quatro das cinco métricas analisadas, o modelo SVM se destaca na especificidade (97,2%), sendo uma escolha viável para aplicações em que a identificação correta de instâncias negativas é crucial. Com excelente pontuação também na precisão (93,9%), acurácia (93,5%), F1 *score* (89,9%), veja a Figura 18.

Esses resultados contribuem para a validação de modelos de AM na classificação de óleos lubrificantes por espectroscopia Raman.

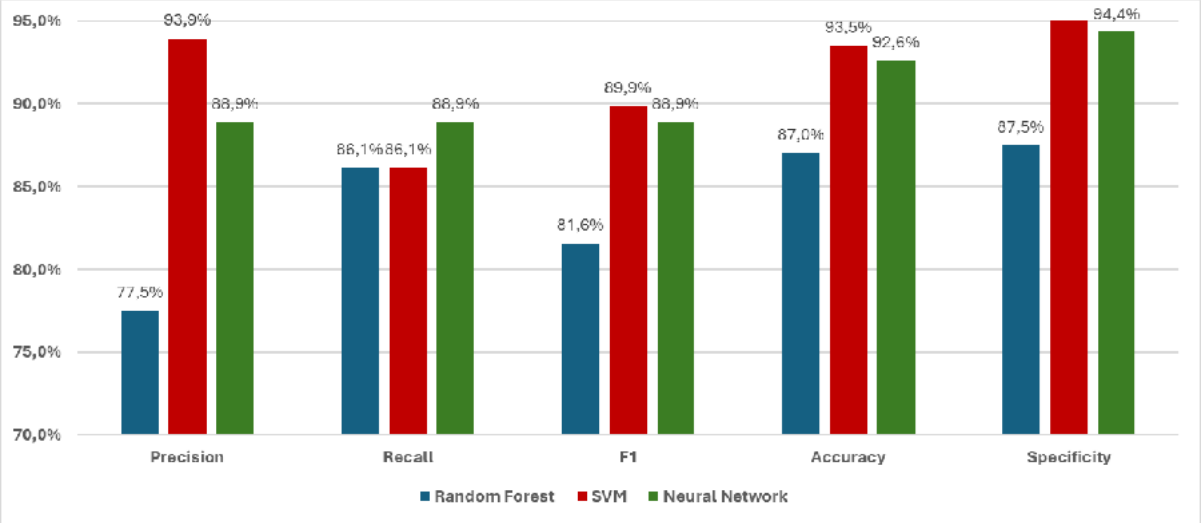

Figura 18 - Comparação entre as métricas de classificação para o óleo semissintético.

3.6.3 Óleo sintético

A partir da matriz de confusão apresentada anteriormente se extrai as métricas essenciais para avaliação dos modelos, como precisão, *recall*, F1-*score*, acurácia e especificidade, veja as Figuras 7, 9, 11.

Observasse na Figura 19 que o SVM e o RN alcançaram 100% de precisão, o que é um desempenho excepcional, ressaltando a confiabilidade desses modelos na identificação correta de óleos lubrificantes sintéticos. Embora o RF também apresentasse uma precisão bem alta (94,3%), indicando que 94,3% das classificações positivas foram corretas.

Sintético	Random Forest	SVM	Neural Network
Precision	94,3%	100%	100%
Recall	91,7%	100%	100%
F1	93,0%	100%	100%
Accuracy	95,4%	100%	100%
Specificity	97,2%	100%	100%

Figura 19 - Comparação entre os modelos de classificação para o óleo sintético.

Os três modelos demonstraram alto *recall*, mas o SVM e a RN atingiram 100%,

indicando que nenhum caso positivo foi negligenciado e 91,7% para o RF, como ser ver na Figura 19. Essa característica é vital na classificação de óleos lubrificantes, pois a omissão de um tipo específico poderia comprometer a eficácia da classificação.

A métrica F1, que combina precisão e *recall* em uma única medida, destaca o equilíbrio entre essas duas dimensões. Novamente o SVM e a RN foram superiores, obtiveram 100%, sugerindo que esses dois modelos não sacrificam precisão em detrimento do *recall* ou vice-versa, proporcionando uma abordagem equilibrada. No entanto, o RF também apresentou um valor alto de F1 de 93%. E o gráfico na Figura 20 da precisão *versus recall é* maior que 90% para todos os modelos analisados, destacando um excelente equilíbrio entre essas duas dimensões. Deste modo os 3 (três) modelos de classificação evidenciaram robustez na classificação de óleos lubrificantes sintéticos, como visto na Figura 19.

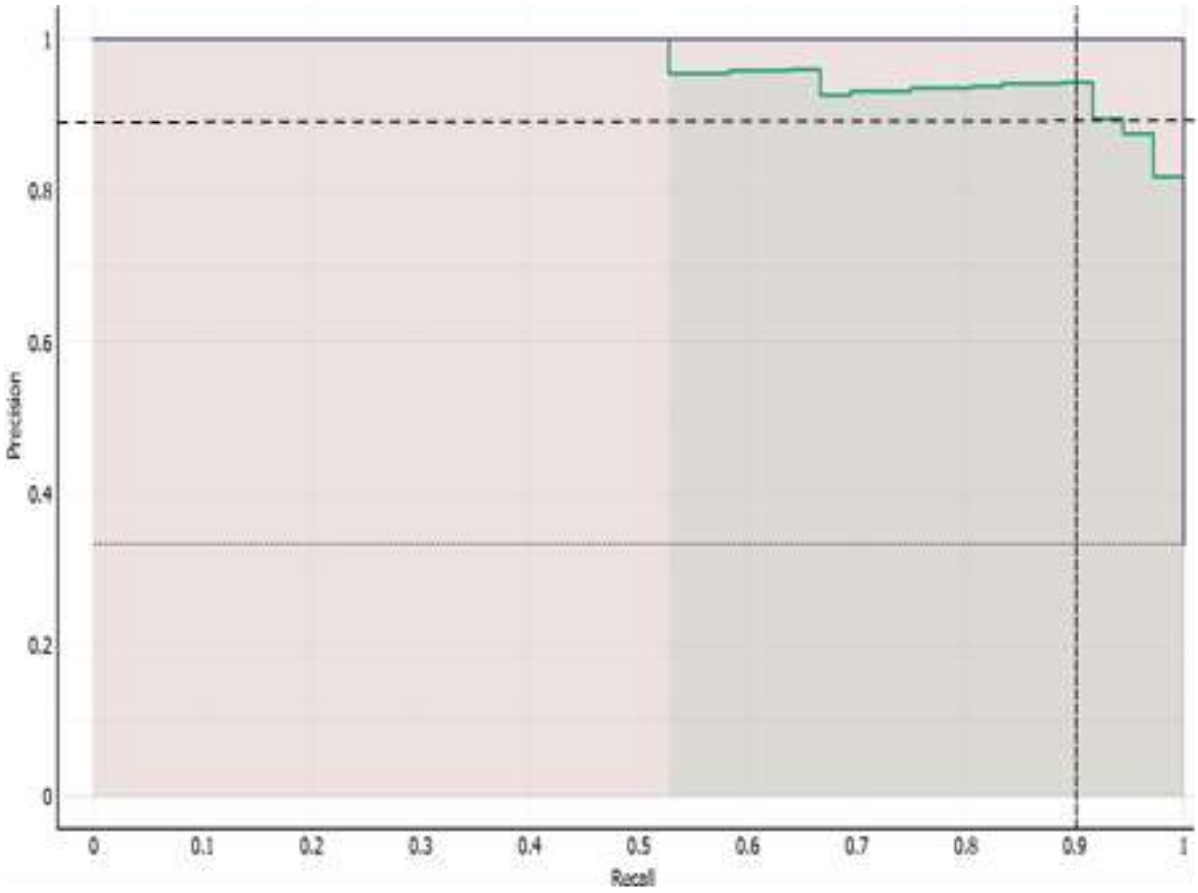

Figura 20 - precisão versus recall, maior que 90% para todos os modelos.

Os resultados na Figura 19, revelam que os algoritmos SVM e a RN alcançaram 100% de acurácia, reforçando a eficácia desses modelos na classificação dos óleos lubrificantes tipo sintético, destacando a confiabilidade dos modelos em uma variedade de situações, embora o modelo RF também tivesse uma alta acurácia de 95,4%.

A especificidade, que mensura a capacidade do modelo em identificar corretamente as instâncias negativas, também é essencial. O SVM e a RN ambos apresentaram 100% de especificidade, respectivamente, veja a Figura 19. Esses valores indicam a habilidade dos modelos citados em evitar falsos positivos, sendo crucial na garantia de que um óleo lubrificante não seja erroneamente classificado como de outro tipo. O desempenho do RF também foi ótimo de 97,2% de especificidade.

A análise dos modelos RF, SVM e a RN apresentam desempenho excepcional na classificação de tipos de óleos lubrificantes sintéticos em motores automotivos a partir de dados de Espectroscopia Raman. No entanto os modelos SVM e a RN foram superiores com 100% em todas as métricas, a saber, precisão, *recall,* F1, acurácia e especificidade, a eficácia e confiabilidade desses modelos, destacam suas aplicabilidades na classificação de óleos lubrificantes, segundo a Figura 21.

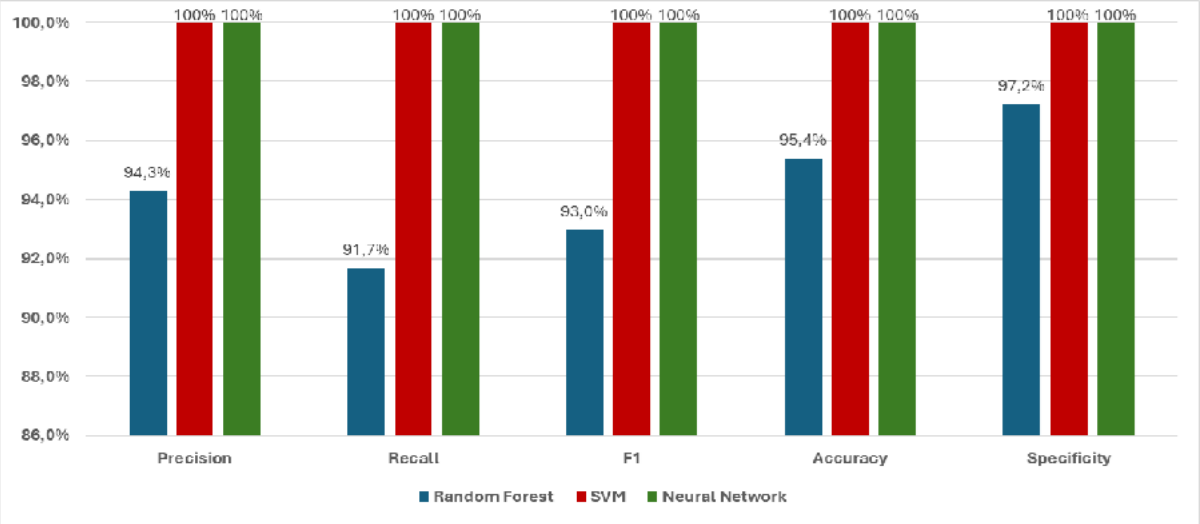

Figura 21 - Comparação entre as métricas de classificação para o óleo sintético.

3.7 Vantagens e desvantagens dos modelos de classificação

A aplicação de modelos de Aprendizado de Máquina (AM) na classificação de tipos de óleos lubrificantes de motores automotivos por meio de espectros Raman oferece uma abordagem promissora para aprimorar a classificação e diferenciação das diversas composições e tipos de óleos lubrificantes automotivos.

No caso do óleo lubrificante mineral, os modelos RF, SVM e RN exibem resultados expressivos em diversas métricas. Enquanto o RF demonstra alta precisão (90,9%) na identificação de óleos minerais, o SVM se destaca pelo melhor recall (94,4%), indicando uma capacidade superior em classificar esses óleos, veja a Figura

13. A análise do F1 *score* revela a harmonia entre precisão e *recall*, com o SVM novamente liderando (90,7%). Embora todos os modelos apresentem acurácia favorável, o SVM se destaca (93,5%). No entanto, o modelo RF se destacou com especificidade (95,8%), evidenciando sua habilidade em distinguir óleos não minerais, com visto na Figura 13.

Para o óleo semissintético, observa-se que o SVM alcança o resultado mais alto em métricas como precisão (93,9%), F1 *score* (89,9%) e acurácia (93,5%), destacando sua capacidade de evitar falsos positivos (FP) e identificar corretamente as instâncias positivas (VP). Além de ter melhor desempenho na métrica especificidade (97,2%), conforme Figura 16. Enquanto o RN se destacou com o maior *recall* (88,9%) também ofereceu desempenho satisfatório nas suas métricas, mostrando um excelente equilíbrio entre precisão e *recall*, veja Figura 16.

No contexto do óleo sintético, os modelos SVM e RN se destacam com resultados excepcionais (100%) em todas as métricas, incluindo precisão, *recall*, F1 *score*, acurácia e especificidade. Todos os modelos estudados demonstram confiabilidade na identificação correta de óleos lubrificantes sintéticos, enfatizando sua aplicabilidade em uma variedade de situações. Enquanto o RF também oferece um desempenho sólido, os modelos SVM e RN mostram-se superiores em todas as métricas, destacando sua eficácia e confiabilidade na classificação de óleos lubrificantes.

3.8 Discussão Final

A obtenção e preparação dos dados de espectroscopia Raman, cedidos pelo Laboratório de Espectroscopia Vibracional do Centro de Inovação, Tecnologia e Educação – CITE da Universidade Anhembi Morumbi, representaram o primeiro passo para a realização deste estudo. Com a utilização de amostras de três tipos distintos de óleos lubrificantes automotivos - mineral 15W30, semissintético 15W40 e sintético 5W30, provenientes da marca LUBRAX da Petrobras Distribuidora S.A., este conjunto de dados proporcionou uma visão detalhada dos diferentes espectros Raman desses produtos multiviscosos, recomendados para uma variedade de motores.

Embora esse conjunto de dados fossem originalmente destinados à identificação da degradação dos óleos sob condições específicas de aquecimento, concentramo-nos, neste estudo, exclusivamente na classificação dessas amostras

quanto ao tipo, a saber em mineral, semissintético e sintético. Passando pelo processo de pré-processamento que incluiu a remoção de ruídos e a normalização dos espectros, garantimos a qualidade e integridade dos dados utilizados em nossos modelos de AM. Com um total de 108 amostras (instâncias) distribuídas em três classes distintas quanto ao tipo de óleo, mineral, semissintético e sintético, sendo cada uma com 36 instâncias, foi possível explorar um conjunto de dados que se mostrou robusto e representativo para treinar e avaliar os modelos de classificação multiclasse.

A implementação e ajuste dos modelos de Random Forest (RF), Support Vector Machine (SVM) e Rede Neural (RN) representaram uma etapa fundamental na consecução do segundo objetivo deste estudo. O processo de pré-processamento teve como objetivo extrair as principais características dos dados, proporcionando uma representação mais eficiente para os modelos de AM. Os modelos RF, SVM e RN foram associados ao widget “Test and score” do software Orange Data Mining para a avaliação do seu desempenho, garantindo uma variedade de abordagens de aprendizado e permitindo uma comparação das suas capacidades em classificar diferentes tipos de óleos lubrificantes.

A escolha desses modelos (RF, SVM. RN) permitiu uma análise representativa dos dados, considerando a variabilidade presente nos dados. Após isso, na fase de teste os resultados foram encaminhados para o widget “Confusion Matrix”, proporcionando uma visão da matriz de confusão para cada modelo implementado, indicando as taxas de verdadeiros positivos, verdadeiros negativos, falsos positivos e falsos negativos. E com a visualização dos resultados em um gráfico de dispersão foi possível uma análise mais detalhada dos casos em que os modelos não conseguiram classificar corretamente os óleos lubrificantes. Como resultado, ao combinar a etapa de preparação dos dados com o processo de treinamento e avaliação, seguido da análise e representação visual dos resultados, teve um papel significativo na avaliação e compreensão do desempenho de modelos de classificação aplicados à espectroscopia Raman em diferentes variedades de óleos lubrificantes automotivos em relação ao estudo.

Avaliar o desempenho dos modelos usando métricas apropriadas, como acurácia, precisão, *recall* e F1-score, foi importante para compreender a eficácia desses algoritmos na classificação de tipos de óleos lubrificantes. A análise detalhada das métricas revelou informações valiosos sobre o comportamento de cada modelo em relação às diferentes classes de óleos. Para o óleo lubrificante **mineral**, os modelos

RF, SVM e RN apresentaram resultados expressivos em diversas métricas. O RF demonstrou alta precisão (90,9%), enquanto o SVM se destacou pelo melhor recall (94,4%), indicando sua capacidade superior em classificar esses óleos. A análise do F1 score revelou harmonia entre precisão e *recall*, com o SVM liderando (90,7%). Embora todos os modelos apresentassem acurácia favorável, o SVM se destacou (93,5%). No entanto, o modelo RF se sobressaiu com especificidade (95,8%), evidenciando sua habilidade em distinguir óleos não minerais. Para o óleo **semissintético**, observou-se que o SVM alcançou os resultados mais altos em métricas como precisão (93,9%), F1 score (89,9%) e acurácia (93,5%), destacando sua capacidade de evitar falsos positivos e identificar corretamente as instâncias positivas. O RN se destacou com o maior *recall* (88,9%), mostrando um excelente equilíbrio entre precisão e *recall*. No contexto do óleo **sintético**, os modelos SVM e RN se destacaram com resultados excepcionais (100%) em todas as métricas, incluindo precisão, *recall*, F1 score, acurácia e especificidade.

Observou-se que o modelo SVM demonstrou consistência em alcançar altas pontuações em todas as métricas avaliadas, destacando-se principalmente pela alta precisão e *recall*, indicando uma capacidade superior em identificar corretamente os óleos lubrificantes. Por outro lado, os modelos RF e RN também apresentaram desempenhos notáveis, cada um com suas próprias vantagens e limitações. Essa avaliação meticulosa forneceu uma base sólida para a comparação e interpretação dos resultados, contribuindo para uma compreensão abrangente do desempenho dos modelos aplicados a dados de espectroscopia Raman.

Comparar os modelos, identificando seus pontos fortes e fracos como desempenho e interpretabilidade, permitiu uma análise mais aprofundada das capacidades e limitações de cada algoritmo. Ao examinar as métricas de desempenho para cada tipo de óleo lubrificante, ficou evidente que cada modelo exibia pontos fortes em métricas específicas, refletindo suas diferentes abordagens de aprendizado. O modelo SVM, por exemplo, destacou-se pela alta precisão e *recall* em todas as classes de óleos, enquanto o RF demonstrou uma excelente especificidade. Por outro lado, a RN mostrou uma boa capacidade de equilibrar precisão e *recall*. Além disso, a interpretabilidade dos modelos também foi considerada, com o RF oferecendo informações visuais detalhados por meio de sua matriz de confusão e o SVM e RN apresentando desempenhos excepcionais em métricas cruciais. Essa análise comparativa permitiu uma avaliação holística dos modelos, identificando suas vantagens e desafios específicos em relação à classificação de óleos lubrificantes.

4 CONCLUSÕES

A análise abrangente dos resultados obtidos destaca a viabilidade do uso de técnicas de Aprendizado de Máquina (AM) na classificação de óleos lubrificantes de motores automotivos com base em espectroscopia Raman. Os modelos Random Forest, Support Vector Machine (SVM) e Neural Network (RN) demonstraram habilidades distintas, cada um com suas próprias vantagens e limitações. Para o óleo lubrificante **mineral**, os modelos RF, SVM e RN apresentaram resultados expressivos em diversas métricas. O RF demonstrou alta precisão (90,9%), enquanto o SVM se destacou pelo melhor recall (94,4%), indicando sua capacidade superior em classificar esses óleos. A análise do F1 score revelou harmonia entre precisão e *recall*, com o SVM liderando (90,7%). Embora todos os modelos apresentassem acurácia favorável, o SVM se destacou (93,5%). No entanto, o modelo RF se sobressaiu com especificidade (95,8%), evidenciando sua habilidade em distinguir óleos não minerais. Para o óleo **semissintético**, observou-se que o SVM alcançou os resultados mais altos em métricas como precisão (93,9%), F1 score (89,9%) e acurácia (93,5%), destacando sua capacidade de evitar falsos positivos e identificar corretamente as instâncias positivas. O RN se destacou com o maior *recall* (88,9%), mostrando um excelente equilíbrio entre precisão e *recall*. No contexto do óleo **sintético**, os modelos SVM e RN se destacaram com resultados excepcionais (100%) em todas as métricas, incluindo precisão, *recall*, F1 score, acurácia e especificidade.

A precisão e a eficiência na identificação dos diferentes tipos de óleos foram notáveis, proporcionando uma base sólida para futuras aplicações práticas na indústria de lubrificantes automotivos. Além disso, a análise comparativa dos modelos permitiu uma compreensão mais profunda de suas capacidades em termos de desempenho e interpretabilidade. A alta precisão e *recall* alcançados pelo SVM em todas as classes de óleos, com a robustez geral demonstrada pelo modelo, destacam sua eficácia na classificação precisa dos óleos lubrificantes. Esses resultados, combinados com a capacidade do RF de oferecer informações visuais detalhados por meio de sua matriz de confusão e a capacidade da RN de equilibrar precisão e *recall*, ressaltam o potencial das técnicas de AM na otimização da identificação e diferenciação de diferentes composições de óleos lubrificantes automotivos.

Em suma, os resultados deste estudo fornecem uma base promissora para o desenvolvimento e implementação de sistemas de classificação automatizados

baseados em espectroscopia Raman na indústria de lubrificantes, contribuindo significativamente para a eficiência e confiabilidade dos processos de manutenção e diagnóstico em veículos automotivos.

4.1Trabalhos futuros

Com base nas análises abordadas para cada tipo de óleo lubrificante, diversas oportunidades para aprimoramento e investigações futuras emergem.

No tocante ao óleo mineral, considera-se explorar abordagens de ajuste de hiperparâmetros para a *Neural Network*, buscando otimizar ainda mais seu equilíbrio entre precisão e *recall*.

Além disso, investigar a inclusão de *features* adicionais provenientes de técnicas avançadas de processamento de dados espectroscópicos poderia aprimorar a capacidade do modelo em capturar nuances específicas desse tipo de óleo.

Para o óleo semissintético, uma análise mais profunda da interpretabilidade dos modelos pode ser conduzida, proporcionando *informações* sobre as características espectrais que influenciam significativamente nas decisões de classificação.

Adicionalmente, explorar a robustez dos modelos diante de variações nas condições experimentais pode contribuir para a generalização dos resultados.

Quanto ao óleo sintético, considera-se investigar a aplicação de técnicas de interpretabilidade, como SHAP (*SHapley Additive exPlanations*), para compreender melhor os fatores determinantes nas decisões dos modelos. A exploração de arquiteturas de redes neurais mais complexas e estratégias de *ensemble* para a *Random Forest* e SVM também apresenta-se como uma vertente interessante para avaliar a maximização do desempenho desses modelos.

Por fim, a validação externa dos modelos em diferentes conjuntos de dados de espectroscopia Raman pode consolidar a robustez das abordagens desenvolvidas, proporcionando uma visão mais abrangente de sua aplicabilidade em diferentes contextos industriais. Esses trabalhos futuros têm o potencial de aprimorar ainda mais a eficácia e aplicabilidade dos modelos na classificação de óleos lubrificantes automotivos.

REFERÊNCIAS

ABRAMCZUK, A. A. **A Prática da tomada de decisão**. São Paulo: Atlas. 2009.

ADHIKARY, K. *et al.* Evaluating the Performance of Various SVM Kernel Functions Based on Basic Features Extracted from KDDCUP'99 Dataset by Random Forest Method for Detecting DDoS Attacks. **WIRELESS PERSONAL COMMUNICATIONS**, v. 123, n. 4, p. 3127 – 3145, April 2022. ISSN 0929-6212.

AMOR, A. B. H.; GHOUL, O. E.; JEMNI, M. Sign Language Recognition Using the Electromyographic Signal: A Systematic Literature Review. **Sensors**, v. 23, n. 19, 2023. Publisher: MDPI. Disponível em: https://www.mdpi.com/1424-8220/23/19/8343.

BELGIU, M.; DRAGUT, L. Random forest in remote sensing: A review of applications and future directions. **ISPRS JOURNAL OF PHOTOGRAMMETRY and REMOTE SENSING**, v. 114, p. 24 – 31, April 2016. ISSN 0924-2716.

BEZERRA, A. **ANÁLISE DA DEGRADAÇÃO DE ÓLEOS LUBRIFICANTES EM FUNÇÃO DA TEMPERATURA ATRAVÉS DA ESPECTROSCOPIA RAMAN**. 2020. 57 p.
Dissertação (Programa de Pós-Graduação em Engenharia Mecânica) — Universidade Santa Cecília. Disponível em: https://unisanta.br/arquivos/mestrado/mecanica/dissertacoes/Dissertacao_ANDRESSACRISTINADEMATTOSBEZERRA429.pdf.

BEZERRA, A. *et al.* Temperature-Induced Chemical Changes in Lubricant Automotive Oils Evaluated Using Raman Spectroscopy. **APPLIED SPECTROSCOPY**, v. 75, n. 2, p. 145 – 155, February 2021. ISSN 0003-7028.

BONESSO, D. Estimação dos parâmetros do kernel em um classificador svm na classificação de imagens hiperespectrais em uma abordagem multiclasse. 2013. Disponível em: https://www.lume.ufrgs.br/handle/10183/86168.

BREIMAN, L. Random forests. **MACHINE LEARNING**, v. 45, n. 1, p. 5 – 32, October 2001. ISSN 0885-6125.

BURGES, C. A tutorial on Support Vector Machines for pattern recognition. **DATA MINING and KNOWLEDGE DISCOVERY**, v. 2, n. 2, p. 121 – 167, June 1998. ISSN 1384-5810.

CANECA, A. *et al.* Assessment of infrared spectroscopy and multivariate techniques for monitoring the service condition of diesel-engine lubricating oils. **TALANTA**, v. 70, n. 2, p. 344 – 352, September 2006. ISSN 0039-9140.

CHERKASSKY, V.; MA, Y. Practical selection of SVM parameters and noise estimation for
SVM regression. **NEURAL NETWORKS**, v. 17, n. 1, p. 113 – 126, January 2004. ISSN 0893-6080.

DEMSAR, J. *et al.* Orange: Data Mining Toolbox in Python. **JOURNAL OF MACHINE LEARNING RESEARCH**, v. 14, p. 2349 – 2353, August 2013. ISSN 1532-4435.

FENG, Z.; ZHENG, L.; LIU, J. Classification of household microplastics using a multi-model approach based on Raman spectroscopy. **CHEMOSPHERE**, v. 325, June 2023. ISSN 0045-6535. Place: THE BOULEVARD, LANGFORD LANE, KIDLINGTON, OXFORD OX5 1GB, ENGLAND Publisher: PERGAMON-ELSEVIER SCIENCE LTD Type: Article.

FERRARI, A.; BASKO, D. Raman spectroscopy as a versatile tool for studying the properties of graphene. **NATURE NANOTECHNOLOGY**, v. 8, n. 4, p. 235 – 246, April 2013. ISSN 1748-3387.

GUAN, L. *et al.* Application of dielectric spectroscopy for engine lubricating oil degradation monitoring. **SENSORS and ACTUATORS A-PHYSICAL**, v. 168, n. 1, p. 22 – 29, July 2011. ISSN 0924-4247.

HAN, J. *et al.* Near-Infrared Spectroscopy Detection of Pollution Concentration of Agricultural Machinery Lubricating Oil Based on Improved Random Frog Algorithm. **SPECTROSCOPY and SPECTRAL ANALYSIS**, v. 42, n. 11, p. 3482 – 3488, November 2022. ISSN 1000-0593.

HE, H.; GARCIA, E. Learning from Imbalanced Data. **IEEE TRANSACTIONS ON KNOWLEDGE and DATA ENGINEERING**, v. 21, n. 9, p. 1263 – 1284, September 2009. ISSN 1041-4347.

HINTON, G.; SALAKHUTDINOV, R. Reducing the dimensionality of data with neural networks. **SCIENCE**, v. 313, n. 5786, p. 504 – 507, July 2006. ISSN 0036-8075.

HUO, W. *et al.* Performance prediction of proton-exchange membrane fuel cell based on convolutional neural network and random forest feature selection. **ENERGY CONVERSION and MANAGEMENT**, v. 243, September 2021. ISSN 0196-8904.

JOLLIFFE, I.; CADIMA, J. Principal component analysis: a review and recent developments.
PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A-MATHEMATICAL PHYSICAL and ENGINEERING SCIENCES, v. 374, n. 2065, April 2016. ISSN 1364-503X.

JORDAN, M.; MITCHELL, T. Machine learning: Trends, perspectives, and prospects. **SCIENCE**, v. 349, n. 6245, p. 255 – 260, July 2015. ISSN 0036-8075.

KUMAR, A. *et al.* Performance optimisation of face recognition based on LBP with SVM and random forest classifier. **INTERNATIONAL JOURNAL OF BIOMETRICS**, v. 15, n. 3-4, p. 389 – 408, 2023. ISSN 1755-8301.

KUPPUSAMY, R.; NIKOLOVSKI, S.; TEEKARAMAN, Y. Review of Machine Learning Techniques for Power Quality Performance Evaluation in Grid-Connected Systems. **SUSTAINABILITY**, v. 15, n. 20, October 2023. ISSN 2071-1050.

LECUN, Y.; BENGIO, Y.; HINTON, G. Deep learning. **NATURE**, v. 521, n. 7553, p. 436 – 444, May 2015. ISSN 0028-0836.

LEI, Y. *et al.* Applications of machine learning to machine fault diagnosis: A review and roadmap. **MECHANICAL SYSTEMS and SIGNAL PROCESSING**, v. 138, April 2020. ISSN 0888-3270.

LIU, J. *et al.* Comparison of Random Forest and Neural Network in Modeling the Performance and Emissions of a Natural Gas Spark Ignition Engine. **JOURNAL OF ENERGY RESOURCES TECHNOLOGY-TRANSACTIONS OF THE ASME**, v. 144, n. 3, March 2022. ISSN 0195-0738.

LIU, W. *et al.* Discrimination of geographical origin of extra virgin olive oils using terahertz spectroscopy combined with chemometrics. **FOOD CHEMISTRY**, v. 251, p. 86 – 92, June 2018. ISSN 0308-8146.

MARRS, J.; NI-MEISTER, W. Machine Learning Techniques for Tree Species Classification Using Co-Registered LiDAR and Hyperspectral Data. **REMOTE SENSING**, v. 11, n. 7, April 2019. ISSN 2072-4292.

OLU-AJAYI, R. *et al.* Data-Driven Tools for Building Energy Consumption Prediction: A Review. **ENERGIES**, v. 16, n. 6, March 2023. ISSN 1996-1073.

PAGANO, T. P. *et al.* Bias and Unfairness in Machine Learning Models: A Systematic Review on Datasets, Tools, Fairness Metrics, and Identification and Mitigation Methods.
Big Data and Cognitive Computing, v. 7, n. 1, March 2023. ISSN 2504-2289. Number: 1 Publisher: Multidisciplinary Digital Publishing Institute. Disponível em: https://www.mdpi.com/2504-2289/7/1/15.

PASSONI, D. **Análise de óleos lubrificantes utilizando dados do espectro Raman**. 2017. 64 p. Dissertação (Programa de Pós-Graduação em Engenharia Mecânica) — Universidade Santa Cecília, Santos, SP,. Disponível em: https: //unisanta.br/arquivos/mestrado/mecanica/dissertacoes/Dissertacao_Douglas.pdf.

PASSONI, D. de J.; PACHECO, M. T. T.; SILVEIRA, L. Raman spectroscopy for the identification of differences in the composition of automobile lubricant oils related to SAE specifications and additives. **Instrumentation Science & Technology**, v. 49, n. 2, p. 164 – 181, August 2020. ISSN 1073-9149. Publisher: Taylor & Francis. Disponível em: https://doi.org/10.1080/10739149.2020.1807356.

PATEL, M. *et al.* Morphology, structure and chemistry of extracted diesel soot-Part I: Transmission electron microscopy, Raman spectroscopy, X-ray photoelectron spectroscopy and synchrotron X-ray diffraction study. **TRIBOLOGY INTERNATIONAL**, v. 52, p. 29 – 39, August 2012. ISSN 0301-679X.

PENG, B. *et al.* Application of Surface-Enhanced Raman Spectroscopy in the Screening of
Pulmonary Adenocarcinoma Nodules. **BIOMED RESEARCH INTERNATIONAL**, v. 2022,
June 2022. ISSN 2314-6133. Place: ADAM HOUSE, 3RD FLR, 1 FITZROY SQ,

LONDON, W1T 5HF, ENGLAND Publisher: HINDAWI LTD Type: Article.

PROBST, P.; WRIGHT, M.; BOULESTEIX, A. Hyperparameters and tuning strategies for random forest. **WILEY INTERDISCIPLINARY REVIEWS-DATA MINING and KNOWLEDGE DISCOVERY**, v. 9, n. 3, May 2019. ISSN 1942-4787.

QURESHI, A. *et al.* Performance evaluation of machine learning models on large dataset of android applications reviews. **MULTIMEDIA TOOLS and APPLICATIONS**, March 2023. ISSN 1380-7501.

SHAO, W. *et al.* The Application of Machine Learning Techniques in Geotechnical Engineering: A Review and Comparison. **MATHEMATICS**, v. 11, n. 18, September 2023. ISSN 2227-7390.

VARGAS-CARDONA, H. *et al.* Artificial intelligence for cervical cancer screening: Scoping review, 2009-2022. **INTERNATIONAL JOURNAL OF GYNECOLOGY & OBSTETRICS**, October 2023. ISSN 0020-7292.

WANG, L. *et al.* Review of machine learning methods applied to enhanced geothermal systems. **ENVIRONMENTAL EARTH SCIENCES**, v. 82, n. 3, February 2023. ISSN 1866-6280.

WANG, S.; CHEN, S. Insights to fracture stimulation design in unconventional reservoirs based on machine learning modeling. **JOURNAL OF PETROLEUM SCIENCE and ENGINEERING**, v. 174, p. 682 – 695, March 2019. ISSN 0920-4105.

XU, J. *et al.* Classification of Lubricating Oil Types Using Mid-Infrared Spectroscopy Combined with Linear Discriminant Analysis-Support Vector Machine Algorithm. **LUBRICANTS**, v. 11, n. 6, June 2023. ISSN 2075-4442.

YAN, C. *et al.* Analysis of handmade paper by Raman spectroscopy combined with machine learning. **JOURNAL OF RAMAN SPECTROSCOPY**, v. 53, n. 2, p. 260 – 271, February 2022. ISSN 0377-0486. Place: 111 RIVER ST, HOBOKEN 07030-5774, NJ USA Publisher: WILEY Type: Article.

ZHOU, M. *et al.* An end-to-end deep learning approach for Raman spectroscopy classification. **JOURNAL OF CHEMOMETRICS**, v. 37, n. 2, February 2023. ISSN 0886-9383. Place: 111 RIVER ST, HOBOKEN 07030-5774, NJ USA Publisher: WILEY Type:Article

APÊNDICE A – PARÂMETROS DE CONFIGURAÇÃO NO SOFTWARE ORANGE

A.1 Parametrização RF

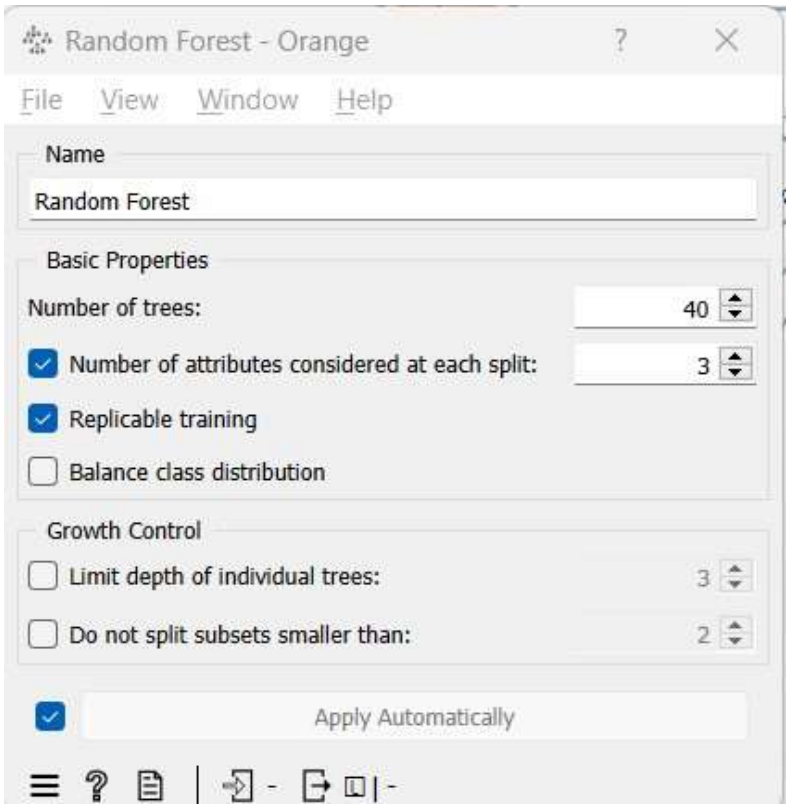

Figura 22 - Parâmetros de configuração do Random Forest

A.2 Parametrização SVM

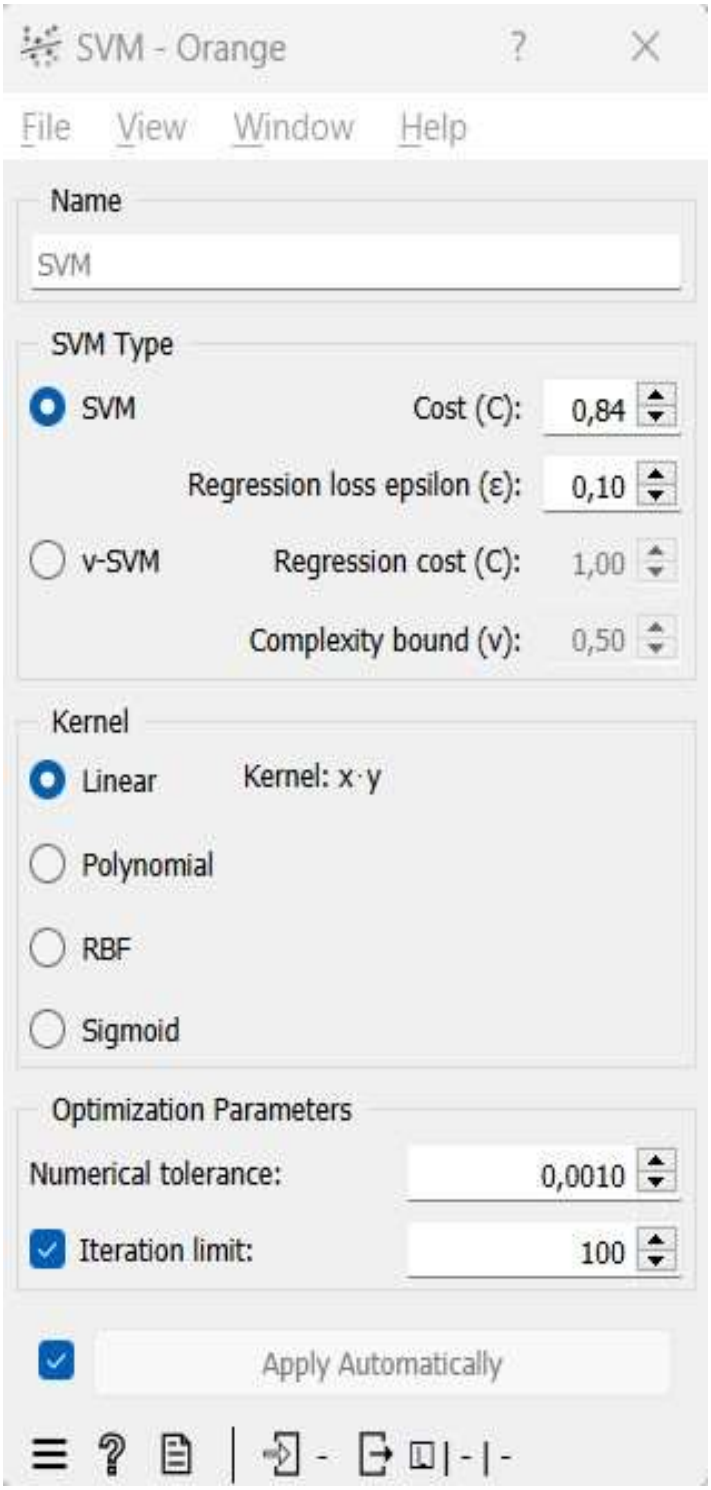

Figura 23 - Parâmetros de configuração do SVM

A.3 Parametrização RN

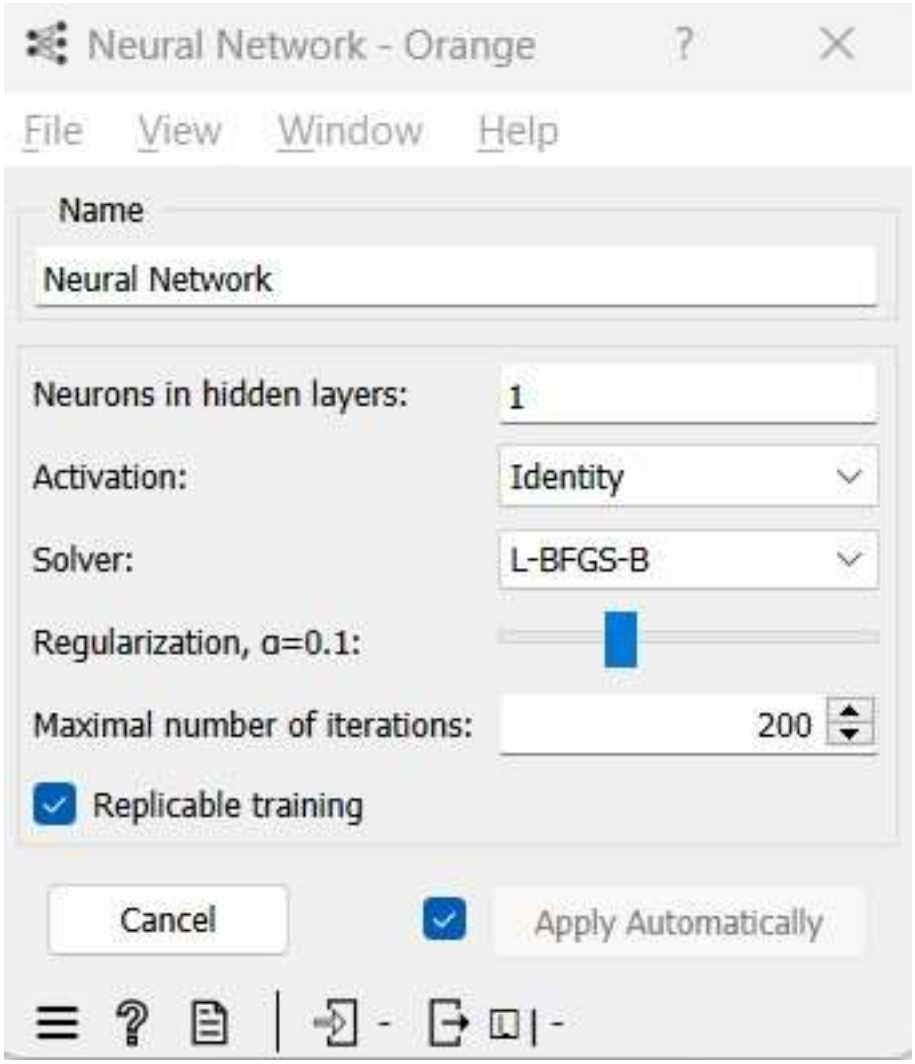

Figura 24 - Parâmetros de configuração da Neural Network

APÊNDICE B – GRÁFICO DE DISPERSÃO DAS COMPONENTES PRINCIPAIS DA PCA

B.1 PC1 e PC2

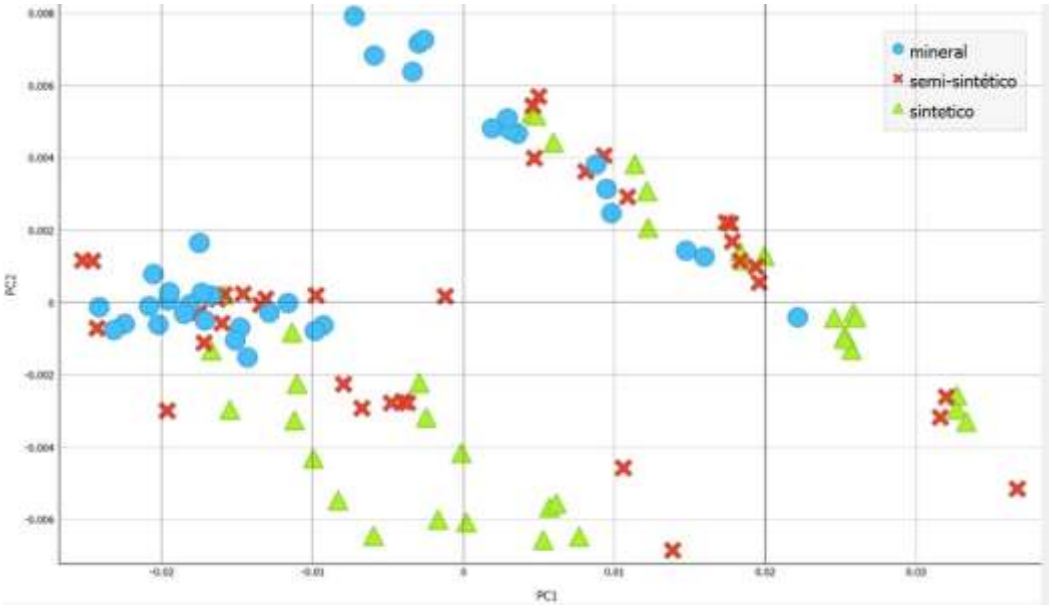

Figura 25 - Projeção de Componentes Principais 1 e 2 (PC1 e PC2)

B.2 PC1 e PC3

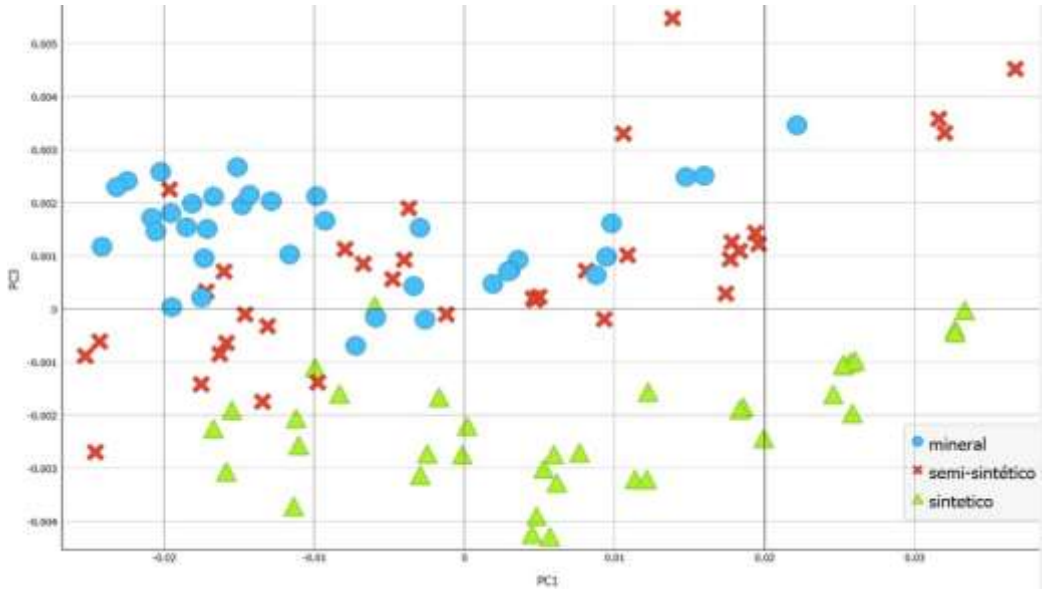

Figura 26 - Projeção de Componentes Principais 1 e 3 (PC1 e PC3)

APÊNDICE C – CURVAS DE VALIDAÇÃO PARA ESCOLHA DOS PARÂMETROS DO MODELO

C.1 Curvas de validação do modelo *Random Forest*

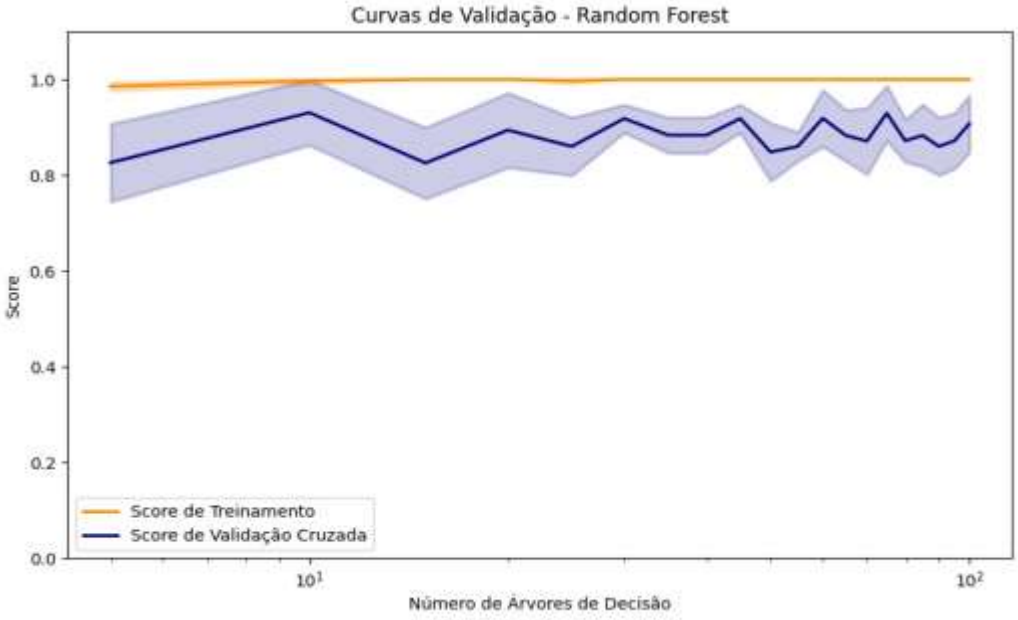

Figura 27 - Curvas de validação do modelo Random Forest.

C.2 Validação dos números de árvores

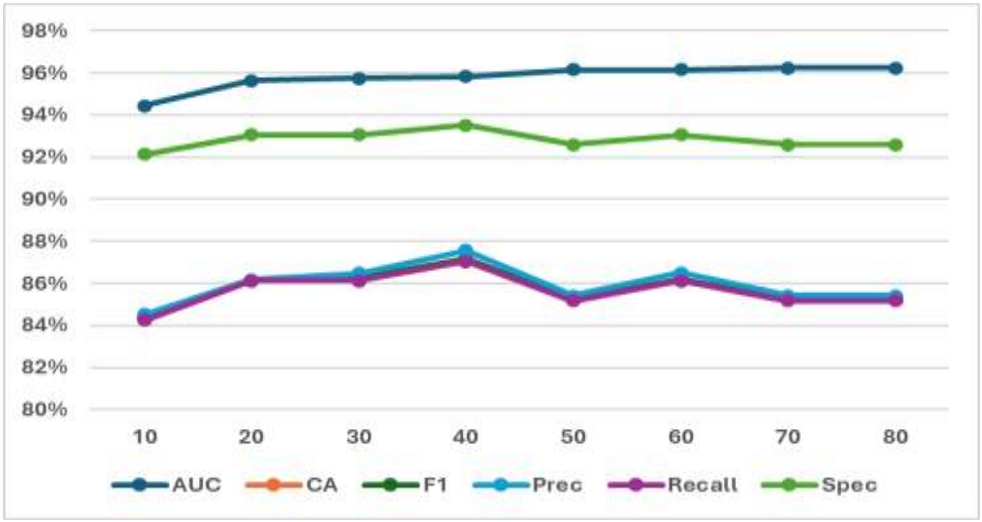

Figura 28 - Validação dos números de árvores

C.3 Curvas de validação do modelo SVM

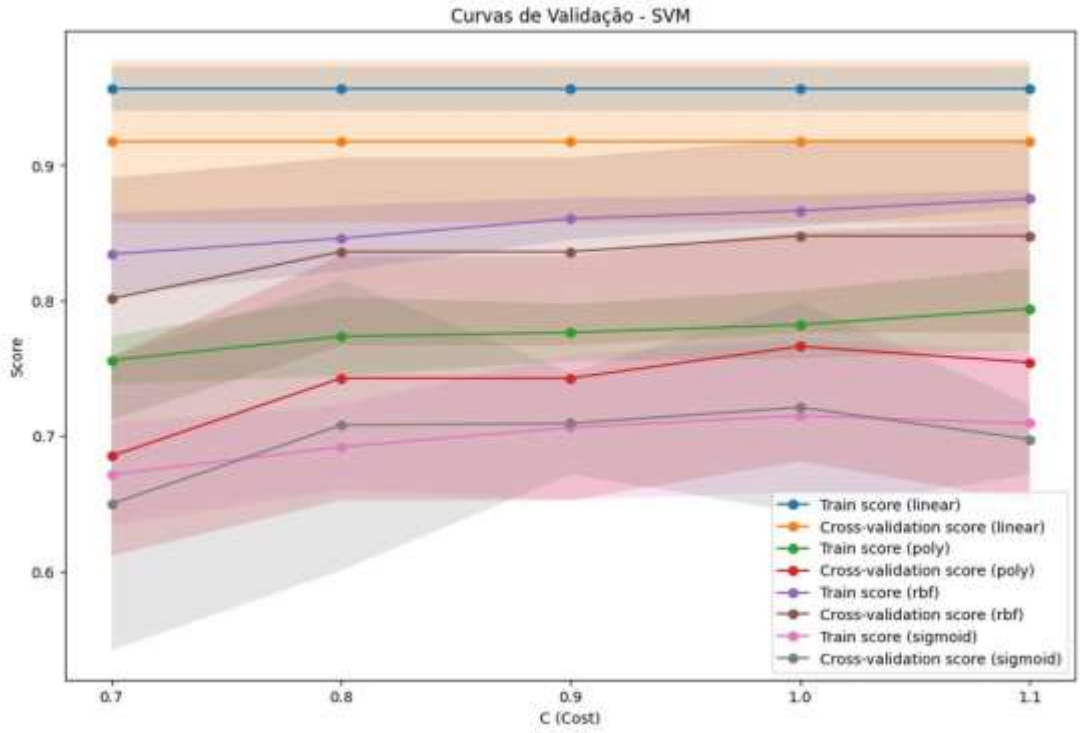

Figura 29 - Curvas de validação do kernel do modelo SVM

C.4 Validação do parâmetro Custo (*Cost*)

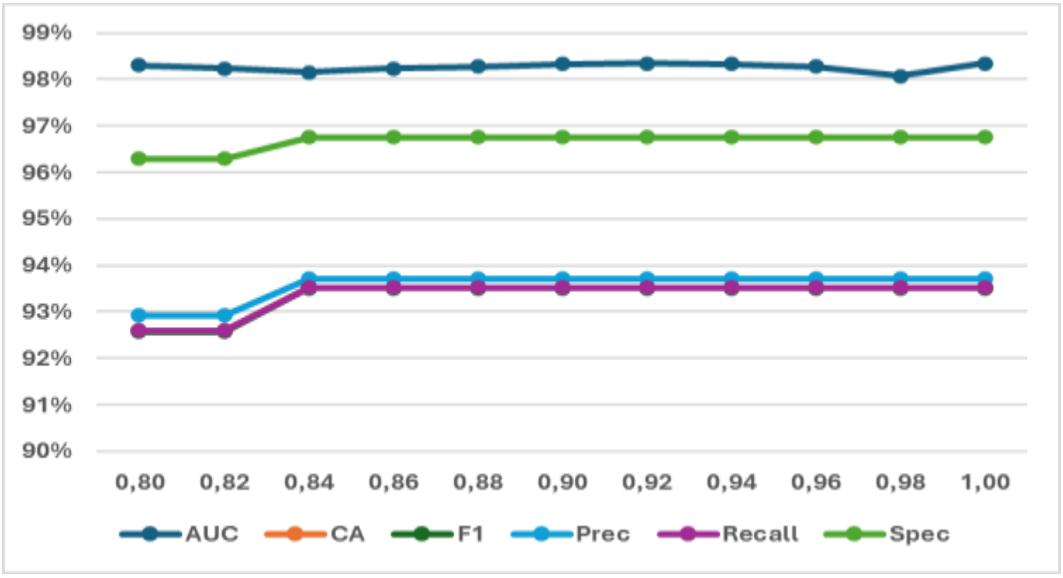

Figura 30 - Validação do parâmetro Custo (Cost)

C.5 Curvas de validação do modelo Neural Network

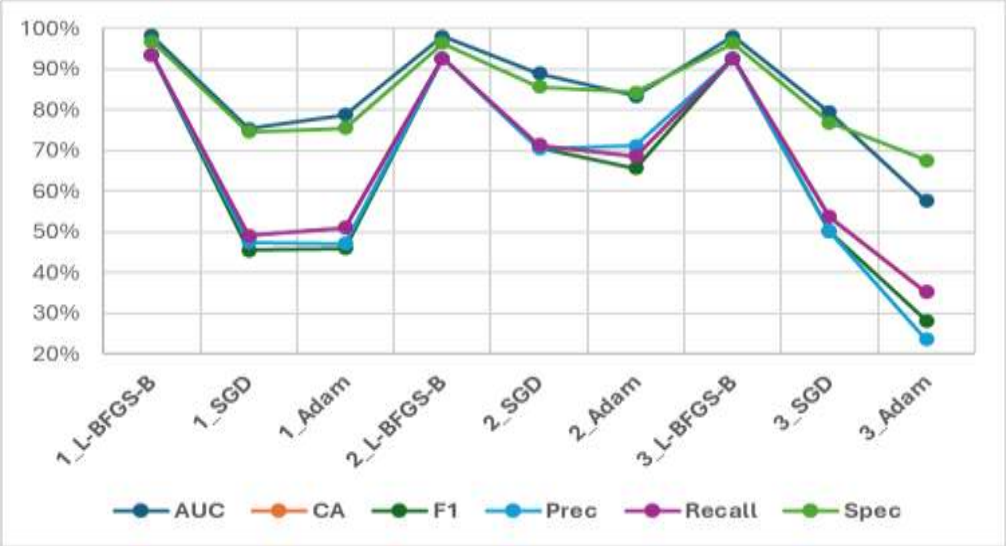

Figura 31 - Número de neurônios na camada oculta e seleção do solver