UNIVERSIDADE FEDERAL DO CEARÁ
CENTRO DE TECNOLOGIA
DEPARTAMENTO DE ENGENHARIA QUÍMICA
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA QUÍMICA

DAVID MATHEUS DE OLIVEIRA ROLIM

DESENVOLVIMENTO DE MODELOS DE APRENDIZADO DE MÁQUINA PARA PREDIÇÃO DE ESTABILIDADE, VISCOSIDADE E TAMANHO MÉDIO DE GOTAS DE EMULSÕES ÁGUA EM ÓLEO

FORTALEZA/CE
2025

DAVID MATHEUS DE OLIVEIRA ROLIM

DESENVOLVIMENTO DE MODELOS DE APRENDIZADO DE MÁQUINA PARA PREDIÇÃO DE ESTABILIDADE, VISCOSIDADE E TAMANHO MÉDIO DE GOTAS DE EMULSÕES ÁGUA EM ÓLEO

Dissertação apresentada ao Programa de Pós-Graduação em Engenharia Química da Universidade Federal do Ceará, como requisito parcial à obtenção do título de Mestre em Engenharia Química. Área de concentração: Processos Químicos e Bioquímicos.

Orientador: Prof. Dr. Filipe Xavier Feitosa.

Coorientadora: Profa. Dra. Andréa da Silva Pereira.

FORTALEZA/CE

2025

Dados Internacionais de Catalogação na Publicação
Universidade Federal do Ceará
Sistema de Bibliotecas
Gerada automaticamente pelo módulo Catalog, mediante os dados fornecidos pelo(a) autor(a)

R653d Rolim, David Matheus de Oliveira.
Desenvolvimento de modelos de aprendizado de máquina para predição de estabilidade, viscosidade e tamanho médio de gotas de emulsões água em óleo / David Matheus de Oliveira Rolim. – 2025.
114 f. : il. color.

Dissertação (mestrado) – Universidade Federal do Ceará, Centro de Tecnologia, Programa de Pós-Graduação em Engenharia Química, Fortaleza, 2025.
Orientação: Prof. Dr. Filipe Xavier Feitosa.
Coorientação: Profa. Dra. Andréa da Silva Pereira.

1. Emulsões água-em-óleo. 2. Aprendizado de Máquina. 3. Propriedades do petróleo. 4. Predição de viscosidade. I. Título.

CDD 660

DAVID MATHEUS DE OLIVEIRA ROLIM

DESENVOLVIMENTO DE MODELOS DE APRENDIZADO DE MÁQUINA PARA PREDIÇÃO DE ESTABILIDADE, VISCOSIDADE E TAMANHO MÉDIO DE GOTAS DE EMULSÕES ÁGUA EM ÓLEO

Dissertação apresentada ao Programa de Pós-Graduação em Engenharia Química da Universidade Federal do Ceará, como requisito parcial à obtenção do título de Mestre em Engenharia Química. Área de concentração: Processos Químicos e Bioquímicos.

Aprovada em: 27 / 06 / 2025.

BANCA EXAMINADORA

_____

Prof. Dr. Filipe Xavier Feitosa (Orientador)
Universidade Federal do Ceará (UFC)

_____

Dra. Rosane Alves Fontes
Petróleo Brasileiro S.A - Petrobras

_____

Prof. Dr. Martín Cismondi Duarte
Universidad Nacional de Córdoba (UNC)

A Ila Maria de Oliveira Rolim

In memoriam

AGRADECIMENTOS

Ao Grupo de Pesquisa em Termofluidodinâmica Aplicada (GPTA) por toda infraestrutura. Ao Professor Dr. Hosiberto Batista de Sant'Ana, por conceder a oportunidade de fazer parte do grupo de pesquisa, por todo apoio, confiança e por todas as sugestões atribuídas durante a minha qualificação.

Um agradecimento especial aos Professores Dr. Filipe Xavier Feitosa e Dra. Andréa da Silva pereira por toda a paciência, disponibilidade, parceria e ensinamentos durante toda a orientação deste trabalho.

Aos amigos do Grupo de Pesquisa em Termofluidodinâmica Aplicada (GPTA), que apesar de vários, não posso deixar de citar o Elton, que dividiu o cansaço diário durante todo a etapa experimental (foram muitas emulsões vividas). Ao meu amigo e camarada Davi, o qual se somasse os tempos de pausa para o café dava uma disciplina de 64horas.

Agradeço aos amigos Moacir, Lucas, Alanderson, Hugo, William e Karol pelos vários ensinamentos e apoio durante essa etapa.

Não poderia deixar de citar os demais membros do grupo: Mauro, Júlia, Letícia, Peterson e Vinicius, cada um a sua maneira tornaram o dia a dia memoráveis.

Um grande agradecimento a Vinicius do GPBio, pela disponibilidade em ajudar na etapa de desenvolvimento dos modelos de IA.

Ao professor Dr. Pedro Felipe Gadelha Silvino, por todas as sugestões atribuídas durante a minha qualificação.

Um agradecimento ao meu grande amigo Valdessandro Farias Dantas, que desde a graduação se fez presente dando vários conselhos e apoio durante essa jornada.

A FUNCAP, pelo apoio financeiro por meio da concessão da bolsa durante todo o mestrado.

“Se fracassar, ao menos que fracasse ousando grandes feitos, de modo que a sua postura não seja nunca a dessas almas frias e tímidas que não conhecem nem a vitória nem a derrota.”
(Theodore Roosevelt)

RESUMO

As emulsões são fundamentais na indústria do petróleo, impactando processos de separação de fases e transporte de fluidos. Este estudo avaliou a eficácia de modelos de aprendizado de máquina (Machine Learning – ML) na previsão da influência de parâmetros-chave sobre a estabilidade de emulsões, a viscosidade aparente e o diâmetro médio das gotas (MDS). Uma metodologia em duas etapas foi empregada. Inicialmente, experimentos com 13 (treze) óleos distintos geraram um conjunto de dados abrangente, variando a fração de água, a temperatura e a concentração de salinidade por meio de um planejamento experimental baseado em Amostragem por Hipercubo Latino (Latin Hypercube Sampling – LHS), para analisar seus efeitos nas propriedades das emulsões. Posteriormente, os dados experimentais foram usados para treinar e testar modelos de AM para classificação (Árvore de Decisão - AD, *Gradient Boosting* - GB, Floresta Aleatória - FA e *Multi-Layer Perceptron* - MLP) e regressão (Regressão Polinomial - RP, *eXtreme Gradient Boosting* - XGBoost, FA e MLP), com os hiperparâmetros otimizados por meio do algoritmo de Otimização por Enxame de Partículas (*Particle Swarm Optimization* – PSO). Foi aplicada a análise SHAP (*SHapley Additive exPlanations*) para quantificar o impacto dos parâmetros de entrada nos resultados do modelo, revelando suas contribuições relativas para as previsões. Os modelos de AM previram com precisão a estabilidade das emulsões, com o modelo GB apresentando a melhor performance com acurácia de 0,938. Para o diâmetro médio de gota, o modelo XGBoost apresentou o melhor desempenho de predição, atingindo R^2 de 0,924. Para a viscosidade, o modelo XGBoost também se sobressaiu ao atingir R^2 de 0,992 nas predições. Os modelos conseguiram capturar padrões complexos de separação de fases, além de demonstrarem fortes correlações entre os valores previstos e experimentais para viscosidade aparente e tamanho das gotas. A análise paramétrica utilizando SHAP revelou que o teor de água (%) foi o fator mais influente na previsão da estabilidade das emulsões, enquanto o °API se destacou como a principal variável para a viscosidade e o diâmetro médio de gota.

Palavras-chave: Emulsões água-em-óleo; Aprendizado de Máquina; Propriedades do petróleo; Predição de viscosidade.

ABSTRACT

Emulsions are essential in the petroleum industry, influencing phase separation and fluid transport processes. This study evaluated the effectiveness of machine learning (ML) models in predicting the influence of key parameters on emulsion stability, apparent viscosity, and mean droplet size (MDS). A two-step methodology was employed. Initially, experiments using thirteen (13) different oils generated a comprehensive dataset by varying water cut, temperature, and salinity concentration through a Latin Hypercube Sampling (LHS) design to analyze their effects on emulsion properties. Subsequently, the experimental data were used to train and test ML models for both classification (Decision Tree – DT, Gradient Boosting – GB, Random Forest – RF, and Multi-Layer Perceptron – MLP) and regression tasks (Polynomial Regression – PR, eXtreme Gradient Boosting – XGBoost, RF, and MLP), with hyperparameters optimized using the Particle Swarm Optimization (PSO) algorithm. SHapley Additive exPlanations (SHAP) analysis was applied to quantify the impact of input parameters on the model outcomes, revealing their relative contributions to the predictions. The ML models accurately predicted emulsion stability, with the GB model achieving the best performance, reaching an accuracy of 0.938. For mean droplet size, the XGBoost model showed the best predictive performance, achieving an R^2 of 0.924. For viscosity, XGBoost also stood out, reaching an R^2 of 0.992. The models were able to capture complex phase separation patterns and demonstrated strong correlations between predicted and experimental values for both apparent viscosity and droplet size. SHAP-based parametric analysis revealed that water content (%) was the most influential factor in predicting emulsion stability, while °API emerged as the key variable for both viscosity and mean droplet size.

Keywords: Water-in-oil emulsions; Machine Learning; Oil properties; Viscosity prediction.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

A/O	Água em óleo
O/A	Óleo em água
O/A/O	Óleo-água-óleo
A/O/A	Água-óleo-água
API	American Petroleum Institute
DMG	Diâmetro médio de gota
GE	Gravidade específica
SARA	Saturados, Aromáticos, Resinas e Asfaltenos
PONA	Parafinas, Olefinas, Naftenos e Aromáticos
PIONA	Parafinas, Isoparafinas, Olefinas, Naftenos e Aromáticos
PINA	Parafinas, Isoparafinas, Naftenos e Aromáticos
MLP	*Multi-Layer Perceptron*
AD	Árvore de Decisão
FA	Floresta Aleatória
XGBoost	eXtreme Gradient Boosting
RP	Regressão Polinomial
AM	Aprendizado de Máquina
IA	Inteligência Artificial
AP	Aprendizado Profundo
LHS	*Latin Hypercube Sampling*
PSO	*Particle Swarm Optmization*
VP	Verdadeiro Positivo
VN	Verdadeiro Negativo
FP	Falso Positivo
FN	Falso Negativo
TVP	Taxa de Verdadeiros Positivos
TFP	Taxa de Falsos Positivos
GI	Ganho de Informação
MP	Mega Pixel
AUC	*Area Under Curve*
ROC	*Receiver Operating Characteristic Curve*

LISTA DE SÍMBOLOS

φ	Fração de água
T	Temperatura
P	Pressão
t	tempo
$\dot{\gamma}$	taxa de cisalhamento
μ_d	Viscosidade da fase dispersa
ρ_d	Densidade da fase dispersa
μ_c	Viscosidade da fase contínua
ρ_c	Densidade da fase dispersa
μ_e	Viscosidade da emulsão
μ_o	Viscosidade do óleo
μ_r	Viscosidade relativa
®	Marca Registrada

SUMÁRIO

1 INTRODUÇÃO

O petróleo bruto raramente é produzido isoladamente, geralmente é acompanhado de água, o que gera diversos problemas durante a sua produção. A água produzida pode se apresentar de duas formas: como água livre (ou seja, que se separa rapidamente) ou como emulsões (Kokal, 2005).

Emulsões são dispersões de um líquido em outro, mais comumente água em óleo ou óleo em água (Sjöblom, 2001). A formação de emulsões é um fenômeno comum no desenvolvimento de campos petrolíferos, ocorrendo devido às altas forças de cisalhamento ao longo do poço, oriundas do uso de bombas, válvulas de estrangulamento, dutos, bem como armazenagem de petróleo bruto e no processamento final.

Durante as diversas etapas de produção e processamento do petróleo, a estabilidade dessas emulsões impacta diretamente a eficiência dos processos, sendo influenciada pela presença de surfactantes naturais, como ácidos naftênicos, asfaltenos e resinas. Além da composição química, fatores como temperatura e pressão tornam a previsão da estabilidade das emulsões uma tarefa complexa (De Oliveira *et al.*, 2018; Lake, 2006).

As propriedades dos fluidos presentes em reservatórios de petróleo são fundamentais para a caracterização do reservatório e desempenham um papel crucial na solução de muitos desafios da engenharia de petróleo, pois dependem diretamente do entendimento dessas características (Khataee; Kasiri, 2010; Khoukhi, 2012).

Embora métodos experimentais sejam comuns para avaliação, nem sempre estão disponíveis devido às limitações financeiras ou técnicas e podem ser demorados. Nesses casos, correlações empíricas derivadas ou modelos de inteligência artificiais podem ser utilizados para prever essas propriedades (Al-Marhoun, 2021; Burke *et al.*, 2024). Ou ainda, como comenta Shakouri *et al*. (2025), desenvolver modelos avançados, como modelos de aprendizado de máquina pode ser útil e prático.

A Inteligência Artificial (IA) e o Aprendizado de Máquina (AM) têm ganhado popularidade em diversos setores, quer no ambiente acadêmico, quer no ambiente industrial. Este fato decorre do reconhecimento do potencial de várias aplicações de IA e AM para automatizar processos enquanto aumentam as capacidades de previsão (Belyadi; Haghighat, 2021). Por esse motivo, atualmente são empregadas técnicas baseadas em inteligência artificial para predição de propriedades como viscosidade, análise de reservatórios, precipitação de asfaltenos, formação de emulsões, identificação de poços e predição de produção (Aïfa, 2014;

Amirian *et al.*, 2015; Cunha *et al.*, 2008; Liu *et al.*, 2023; Nandi *et al.*, 2010; Santos *et al.*, 2025; Silva *et al.*, 2021; Talebi *et al.*, 2014).

Nesse contexto, o presente trabalho pretende avaliar como fatores como salinidade, temperatura e cortes de água influenciam as propriedades de emulsões de petróleo, como a estabilidade, a viscosidade aparente e o tamanho médio de gota. Para isso, serão empregados modelos de inteligência artificial, capazes de analisar grandes volumes de dados e identificar padrões complexos, contribuindo para uma compreensão mais aprofundada da formação e estabilidade das emulsões, além de fornecer informações relevantes para a melhoria da eficiência dos processos na indústria petrolífera.

1.1 OBJETIVO GERAL

- Desenvolver modelos baseados em inteligência artificial e aprendizado de máquina para predizer a estabilidade, a viscosidade aparente e o tamanho médio de gotas em emulsões de petróleo do tipo água em óleo (A/O).

1.2 OBJETIVOS ESPECÍFICOS

- Estabelecer uma metodologia para obtenção de emulsões do tipo água-em-óleo (A/O), variando parâmetros como tipo de óleo, temperatura, teor de água e salinidade, de modo a construir um banco de dados destinado ao treinamento e à validação de modelos de Inteligência Artificial.

- Aplicar a amostragem por hipercubo latino (*Latin Hypercube Sampling* – LHS) para otimizar o planejamento experimental, reduzindo o número de experimentos necessários e permitindo avaliar a influência das variáveis de processo nas características das emulsões.

- Desenvolver um algoritmo baseado na Transformada de Hough para círculos (HoughCircles), com o objetivo de identificar e quantificar gotas de água em fotomicrografias, determinando o tamanho médio das gotas de forma automatizada.

- Realizar análises paramétricas das variáveis de entrada, a fim de verificar a magnitude de seus impactos sobre as predições geradas pelos modelos de Aprendizado de Máquina.

2 REVISÃO BIBLIOGRÁFICA

Nesta seção serão apresentados os princípios teóricos a respeito da química de petróleo, das emulsões petrolíferas, da inteligência artificial e do aprendizado de máquina. Além disso, serão tratados dos fatores que alteram a estabilidade de sistemas emulsionados do tipo água em óleo.

2.1 DEFINIÇÃO E QUÍMICA DO PETRÓLEO

O petróleo, também conhecido como óleo bruto, é uma mistura natural complexa composta predominantemente por hidrocarbonetos (em mais de 90% de sua composição). Ele também contém compostos orgânicos derivados de enxofre, nitrogênio, oxigênio e organometálicos (Abdel-Raouf, 2012; Speight, 1997). Os hidrocarbonetos, formados por hidrogênio e carbono, exibem grande variação em sua estrutura molecular, abrangendo desde o metano, que compõe o gás natural, passando por líquidos refinados em gasolina, até ceras cristalinas (Carreón *et al.*, 2021; Speight, 2014).

Diversas classificações dos compostos do petróleo são relatadas na literatura, que levam em consideração as propriedades do óleo bruto, as propriedades dos destilados, a estrutura química, a origem, dentre outras. Entre as classificações mais comuns, referenciadas por seus acrônimos, estão a PONA, PIONA, PINA e SARA:

- PONA: Parafinas, Olefinas, Naftenos e Aromáticos;
- PIONA: Parafinas, Isoparafinas, Olefinas, Naftenos e Aromáticos;
- PINA: Parafinas, Isoparafinas, Naftenos e Aromáticos;
- SARA: Saturados, Aromáticos, Resinas e Asfaltenos.

Das classificações acima citadas, a mais comumente utilizada para classificar os hidrocarbonetos do petróleo é a classificação SARA (Speight, 2014). Uma vez que os saturados englobam os alcanos e cicloparafinas; aromáticos representam hidrocarbonetos mono, di e poliaromáticos; as resinas englobam as frações compostas por moléculas polares contendo heteroátomos como nitrogênio, oxigênio ou enxofre; e os asfaltenos são semelhantes às resinas, mas caracterizados por um maior peso molecular e um núcleo poliaromático (Abdel-Raouf, 2012).

Além da classificação quanto a constituição do petróleo, operacionalmente (para fins de transporte), a viscosidade e a densidade do óleo cru (ou gravidade API) são os parâmetros usualmente utilizados (Ancheyta, 2013). A densidade e a gravidade API ou $°API$ correlacionam-se de acordo a Equação (1), em que GE representa a gravidade específica do óleo a 60 °F. A Tabela 1 apresenta a classificação do petróleo de acordo com o °API.

$$°API = \frac{141}{GE} - 131 \tag{1}$$

Tabela 1 - Classificação de petróleos segundo a Agência Nacional do Petróleo, Gás Natural e Biocombustíveis

Classificação	**°API**
Leve	31 ≤ °API
Médio	22 ≤ °API < 31
Pesado	°API < 22

Fonte: Agência Nacional do Petróleo, Gás Natural e Biocombustíveis (2024)

2.2 EMULSÕES

Uma emulsão é uma mistura na qual um líquido é disperso em outro líquido com o qual não é solúvel, formando pequenas gotas. O líquido presente na forma de gotas é chamado de fase dispersa, enquanto o líquido que sustenta essas gotas, mantendo-as suspensas, é denominado fase contínua (Lake, 2006).

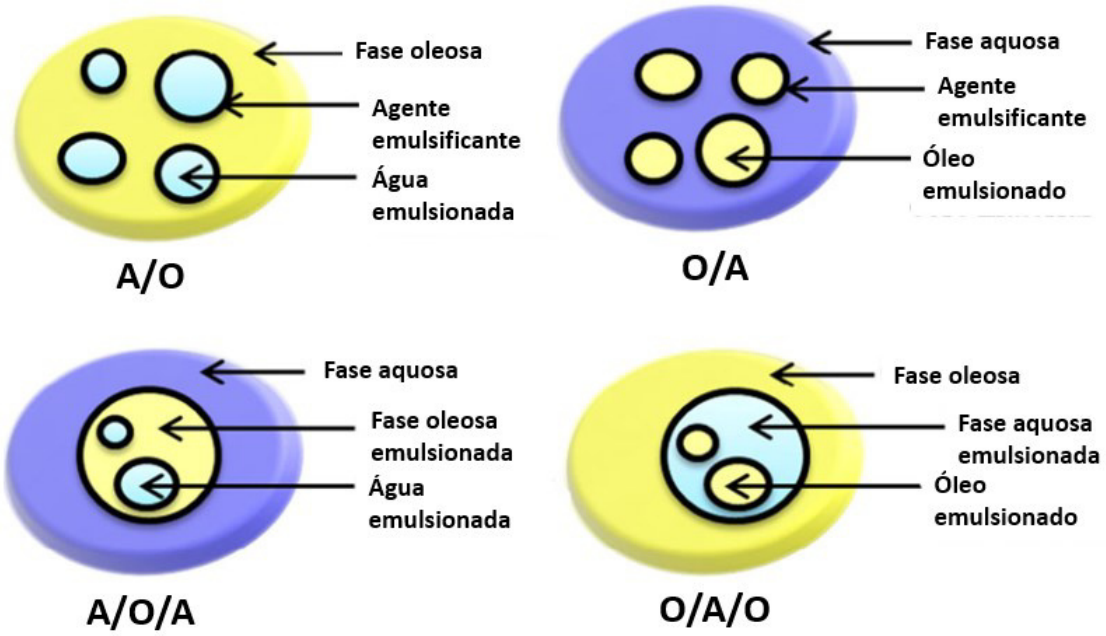
Emulsões estáveis de água e óleo podem se formar durante a produção de óleo bruto em campos petrolíferos, à medida que a água coproduzida é misturada com o óleo, desde o reservatório até as instalações de separação. A depender da unidade de produção, a proporção de água pode variar, podendo ser inferior a 1% ou até ultrapassar 80% em alguns casos (Lake, 2006; Sjöblom *et al.*, 2003).

Emulsões apresentam-se como um problema na produção de petróleo, frequentemente consideradas indesejáveis e podendo acarretar elevados custos de bombeamento, corrosão das tubulações, diminuição da capacidade de transporte e exigência de equipamentos especiais para manuseio (Abdel-Raouf, 2012).

2.2.1 Tipos de emulsões

Em unidades de produção de petróleo, as emulsões são divididas em três categorias principais, a saber: água em óleo (A/O), óleo em água (O/A) e emulsões múltiplas ou complexas. Nas emulsões A/O, pequenas gotas de água são dispersas em uma fase contínua de óleo. Já nas emulsões O/A, são as gotas de óleo que se dispersam em uma fase contínua de água. As emulsões múltiplas ou complexas contêm gotas menores suspensas dentro de gotas maiores, que, por sua vez, estão suspensas em uma fase contínua (Abdulredha; Siti Aslina; Luqman, 2020; Lake, 2006). A Figura 1 ilustra as conformações que as emulsões podem assumir.

Figura 1 - Ilustração dos tipos de emulsões.



Fonte: Adaptação de Bakry *et al.* (2016).

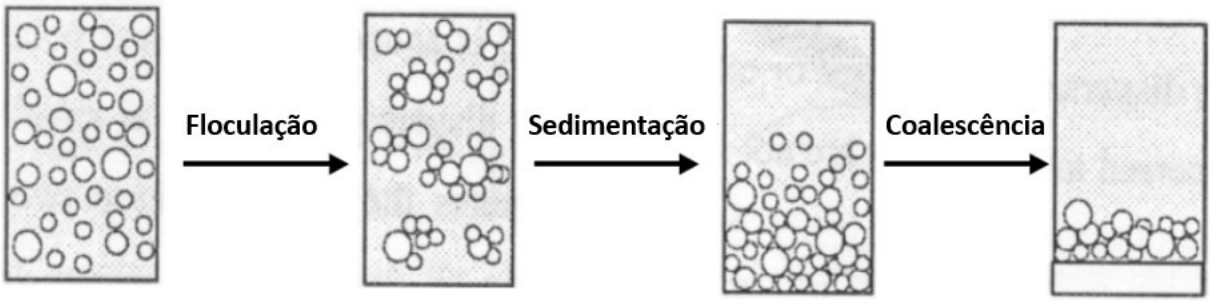
2.2.2 Estabilidade de emulsões

Termodinamicamente, as emulsões são sistemas instáveis que tendem a se separar. Contudo, muitas emulsões demonstram estabilidade cinética, o que permite que permaneçam emulsionadas por longos períodos. Quanto mais prolongado for esse período, maior será a estabilidade da emulsão. Essa estabilidade cinética ocorre devido à formação pequenas gotas (gotas menores normalmente tornam a emulsão mais estável, implicando em um tempo de separação maior) e à criação de um recobrimento interfacial por agentes emulsificantes ou

estabilizantes (que podem ocorrer naturalmente no petróleo bruto ou serem adicionados durante a produção). Estes fatores acabam por suprimir os mecanismos envolvidos na quebra da emulsão (floculação ou agregação, sedimentação, coalescência e inversão de fase)(Abdel-Raouf, 2012; Kokal, 2005; Lake, 2006).

O primeiro passo na desestabilização de emulsões é a floculação, ocorrendo a aglomeração das gotas de água sem variação na área superficial. A sedimentação é a parte do processo que envolve o assentamento por diferença de densidade e a coalescência, por fim, implica a fusão das gotas para formar gotas maiores, até finalmente ocorrer a separação de fases (Lake, 2006; Umar *et al.*, 2018). A representação visual destes mecanismos é apresentada na Figura 2.

Figura 2 - Representação dos mecanismos de separação de emulsões.



Fonte: Adaptação de Abdel-Raouf (2012).

De acordo com Kokal (2005), a classificação das emulsões produzidas em campos de petróleo com base em seu grau de estabilidade cinética são:

- **Emulsões fracas:** Óleo e água separam-se em poucos minutos. A água separada às vezes é chamada de água livre.

- **Emulsões médias:** Separam-se em dez minutos ou mais.

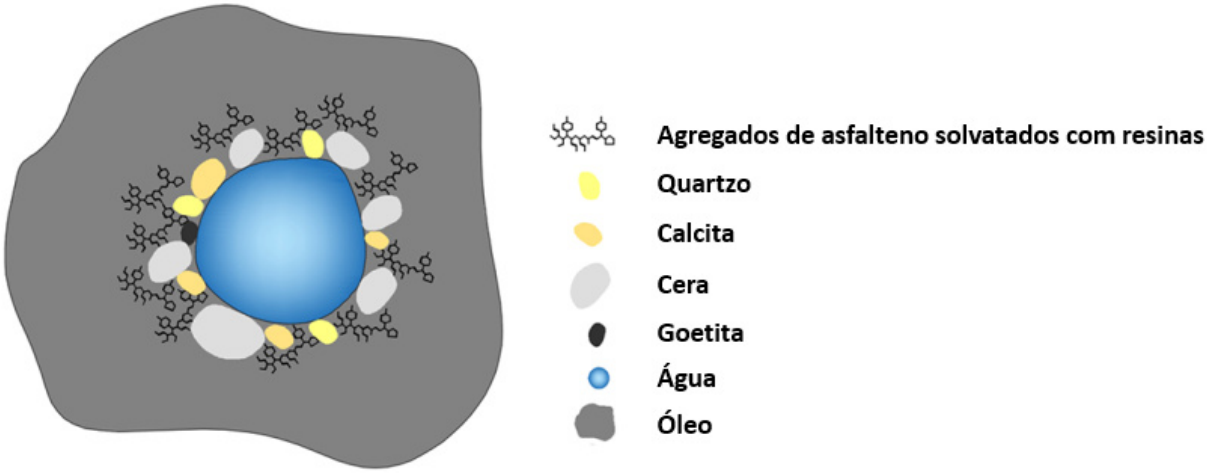
- **Emulsões fortes:** Separam-se, às vezes apenas parcialmente, em algumas horas ou até dias.

A estabilidade da emulsão pode ser determinada pelo tipo e quantidade de agentes ativos de superfície ou surfactantes que podem ocorrer naturalmente no petróleo bruto. Os compostos emulsificantes naturais do petróleo estão nas frações pesadas, como asfaltenos,

resinas, ácidos orgânicos e bases. Esses surfactantes tendem a se concentrar na interface água/óleo, onde formam filmes interfaciais, estabilizando a emulsão ao reduzir a tensão interfacial e promovendo a emulsificação e dispersão das gotas (Abdel-Raouf, 2012; Tchoukov *et al.*, 2014).

Além dos agentes surfactantes, algumas partículas sólidas finas presentes no petróleo bruto são capazes de estabilizar emulsões por difusão na interface óleo/água, formando estruturas rígidas que podem impedir a coalescência das gotas. Para agir como estabilizadores na emulsão, as partículas devem ser muito menores que o tamanho das gotas da emulsão, além de possuir afinidade tanto à fase aquosa quanto à fase oleosa. Entre as partículas que estabilizam a emulsão estão os compostos inorgânicos, como $CaCO_3$ e $CaSO_4$, além de argila, areia, asfaltenos e parafinas, produtos de corrosão, incrustações minerais e lamas de perfuração (Abdel-Raouf, 2012; Lake, 2006; Umar *et al.*, 2018). A Figura 3 ilustra a estabilização de gotas de água por componentes presentes no petróleo.

Figura 3 - Representação esquemática do efeito estabilizador de asfalteno, resinas e outros sólidos em uma gota de água.



Fonte: Adaptação de Sousa *et al*. (2022).

Além disso, outras características que afetam a estabilidade das emulsões são a temperatura e salinidade da salmoura. A temperatura pode atuar afetando a estabilidade de emulsões modificando as propriedades físicas do óleo, da água, dos filmes interfaciais e a solubilidade dos surfactantes nas fases oleosa e aquosa. Além disso, aumenta a energia térmica das gotas e, consequentemente, aumenta a frequência das colisões entre elas. O aumento da temperatura leva a uma desestabilização gradual dos filmes interfaciais (Abdel-Raouf, 2012).

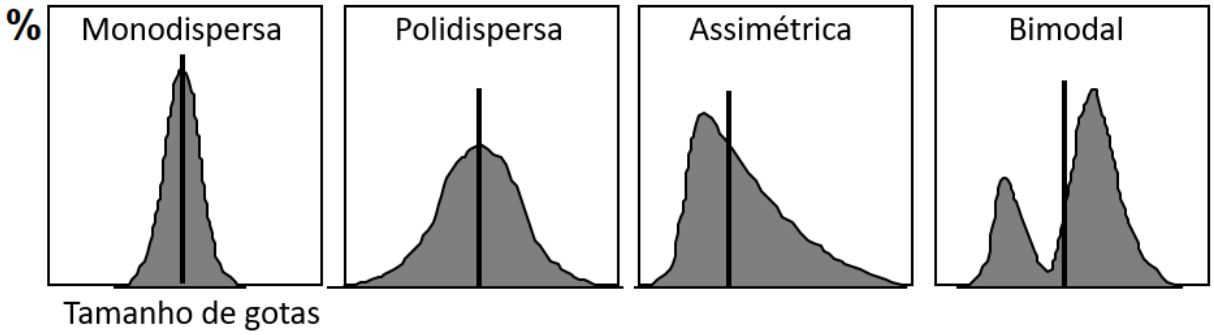
Por sua vez, a salmoura contém como solutos mais comuns os cloretos de sódio (NaCl) e de cálcio ($CaCl_2$), embora outros compostos estejam presentes em menores concentrações. A composição e a salinidade da salmoura impactam a estabilidade da emulsão, influenciando a distribuição do tamanho das gotas. Emulsões do tipo água-em-óleo com menor concentração de NaCl tendem a ser mais estáveis do que aquelas com concentrações mais elevadas (Sousa; Matos; Pereira, 2022).

2.2.3 Caracterização das emulsões

2.2.3.1 Distribuição de tamanho de gotas

De acordo com Salager (2001), pode-se dizer que a distribuição do tamanho das gotas é a impressão digital da emulsão. Uma emulsão é caracterizada principalmente pelo tamanho de gotas, ou mais exatamente, pela distribuição estatística do tamanho das gotas. Dependendo do modo como a emulsão é produzida, particularmente das condições mecânicas de fluido em que o cisalhamento ou a turbulência produziram as gotas, a emulsão deve conter gotas de tamanhos semelhantes ou muito diferentes, com a variedade associada na distribuição estatística. A Figura 4 mostra diferentes formatos das distribuições estatísticas de tamanho de gotas.

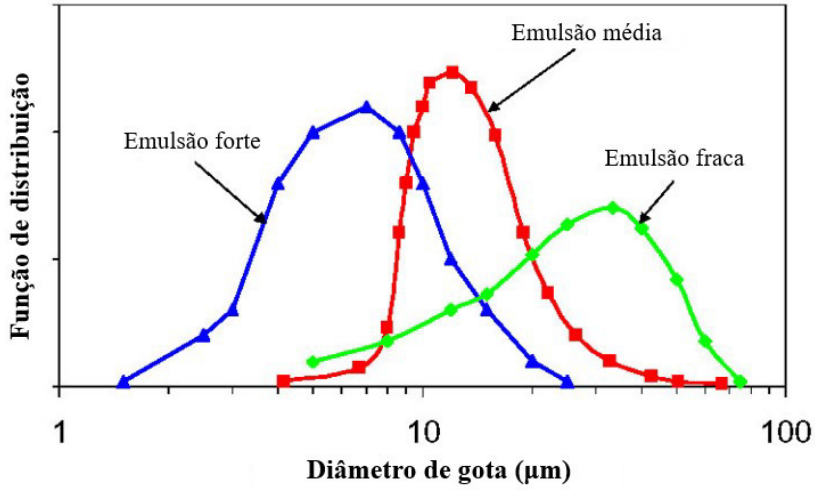
Figura 4 - Diferentes formatos de distribuição de tamanho de gotas.



Fonte: Adaptação de Salager *et al.* (2001).

Entre os fatores que afetam a forma da distribuição do tamanho das gotas em uma emulsão, estão incluídas a tensão interfacial, o cisalhamento, a natureza e a quantidade de agentes emulsificantes, presença de sólidos e propriedades volumétricas do óleo e da água (Lake, 2006). A Figura 5 ilustra diferentes distribuições de tamanho de gotas em relação ao grau de estabilidade cinética.

Figura 5 - Classificação da estabilidade de emulsões de acordo com a distribuição de tamanho de gota.



Fonte: Adaptação de Kokal (2002).

Já que a maioria das emulsões são produzidas por um processo de agitação que muitas vezes envolve turbulência e, portanto, efeitos aleatórios, as emulsões produzidas em campos petrolíferos geralmente têm diâmetros de gotas que variam de 0,1 µm a mais de 100 µm, com seus tamanhos geralmente representados por uma função de distribuição (Lake, 2006; Salager; Isabel Briceno; Luis Bracho, 2001).

Há diversos métodos para se determinar a distribuição do tamanho das gotas para emulsões de campos petrolíferos, entre eles, Lake (2006) destaca:

- propriedades elétricas, como condutividade e constantes dielétricas;
- espalhamento, como espalhamento de luz, espalhamento de nêutrons e espalhamento de raios-X. Essas técnicas cobrem tamanhos de gotas de 0,4 nm a mais de 100 µm;
- separação física, incluindo técnicas cromatográficas, técnicas de sedimentação e fracionamento por fluxo de campo;
- microscopia e análise de imagem.

A distribuição do tamanho das gotas em uma emulsão determina, em certa medida, a estabilidade da emulsão e deve ser considerada na seleção dos protocolos de tratamento ideais. Como regra geral, quanto menor o tamanho médio das gotas de água dispersas, mais apertada

é a emulsão e, portanto, maior o tempo de residência necessário em um separador, o que implica tamanhos maiores de equipamentos de planta de separação (Lake, 2006).

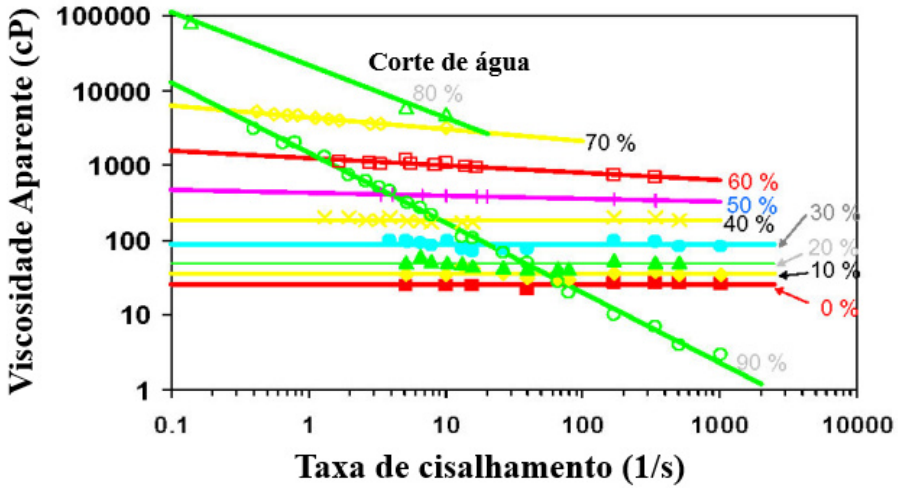
2.2.3.2 Viscosidade

A viscosidade é uma propriedade extremamente importante na exploração, no transporte e no processamento de fluidos petrolíferos. Tanto a extração quanto o transporte são significativamente influenciadas por ela, sendo o fluxo através de meios porosos ou seções de área reduzida fortemente dependente dessa propriedade física (Carreón *et al.*, 2021).

Tratando-se de emulsões, a viscosidade ganha mais relevância uma vez que essa propriedade pode ser consideravelmente maior do que a da fase contínua ou da fase dispersa quando isoladas (Lake, 2006; Salager; Isabel Briceno; Luis Bracho, 2001).

A viscosidade de uma emulsão é influenciada por uma combinação de fatores, incluindo a fração de água, a distribuição do tamanho das gotas, a quantidade de sólidos presentes, as viscosidades do óleo e da água, a taxa de cisalhamento e a temperatura (Sousa; Matos; Pereira, 2022). A Figura 6 ilustra o comportamento da viscosidade aparente de emulsões em função da taxa de cisalhamento para diferentes cortes de água na temperatura de 51,67 °C.

Figura 6 - Viscosidade aparente de emulsões em função do corte de água a uma mesma temperatura.



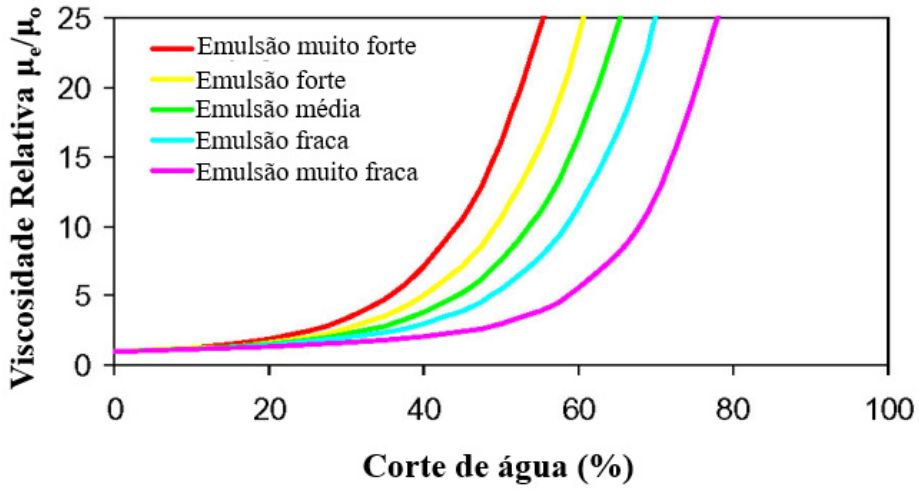
Fonte: Adaptação de Kokal (2002).

Segundo explica Kokal (2002), os dados de viscosidade da Figura 6 mostram que a emulsão apresenta comportamento Newtoniano até um teor de água de 30%, com viscosidade constante para todas as taxas de cisalhamento. Acima de 30%, a emulsão exibe comportamento não newtoniano, caracterizado por comportamento pseudoplástico, no qual a viscosidade

diminui com o aumento das taxas de cisalhamento. As viscosidades aumentam significativamente com teores de água até 80%, muito superiores às da água e do óleo (aproximadamente 1 e 80 cP, respectivamente). Em cerca de 80% de teor de água, a emulsão A/O inverte para O/A, com a água passando de fase dispersa para contínua. Em teores de água muito altos (>95%), foram observadas emulsões múltiplas do tipo água-em-óleo-em-água.

Uma forma de apresentar a viscosidade de emulsões é na forma de viscosidade relativa, em que esta é a razão entre a viscosidade da emulsão e a do óleo na mesma temperatura. A Figura 7 mostra o comportamento da viscosidade relativa em função do corte de água.

Figura 7 - Viscosidade relativa em função do corte de água a uma mesma temperatura.



Fonte: Adaptação de Kokal (2002).

A distribuição do tamanho das gotas afeta a viscosidade da emulsão, sendo maior quando as gotas são menores. A viscosidade da emulsão também será maior quando a distribuição do tamanho das gotas for estreita, ou seja, o tamanho das gotas for relativamente constante (Abdel-Raouf, 2012).

A temperatura também estabelece efeitos em relação à viscosidade da emulsão, estando relacionadas de forma inversamente proporcional. Quando a temperatura aumenta, observa-se uma diminuição na viscosidade da emulsão, causada principalmente pela diminuição da viscosidade do óleo (Abdel-Raouf, 2012; Sousa; Matos; Pereira, 2022).

Estudos indicam que outro fator que contribui para alterações na viscosidade é a salinidade, em que o aumento da salinidade implica em aumento dessa propriedade física (Azodi; Solaimany Nazar, 2013).

2.3 MODELOS NA LITERATURA PARA A DETERMINAÇÃO DA VISCOSIDADE DE EMULSÕES

A viscosidade de emulsões A/O é influenciada por uma série de fatores: temperatura (T), pressão (P), taxa de cisalhamento ($\dot{\gamma}$), fração volumétrica (φ), viscosidade (μ_d) e densidade (ρ_d) da fase dispersa, viscosidade (μ_c) e densidade (ρ_c) da fase contínua e o tamanho médio das gotas. Esses elementos interagem de forma complexa, influenciando uns aos outros. Por exemplo, diferentes taxas de cisalhamento podem alterar o tamanho das gotas, enquanto o aumento da temperatura reduz a viscosidade tanto da fase dispersa quanto da contínua. Essa complexidade torna desafiador reunir todos esses fatores em um único modelo de cálculo (Li *et al.*, 2016).

Contudo, na pesquisa sobre emulsões A/O, diversos estudiosos propuseram fórmulas para calcular a viscosidade relativa (μ_r). Essa viscosidade é descrita como a relação entre a viscosidade da emulsão (μ_e) e a viscosidade do óleo (μ_o), como mostra a Equação (2).

$$\mu_r = \frac{\mu_e}{\mu_o} \tag{2}$$

Várias equações e correlações de viscosidade tratam exclusivamente da fração volumétrica da fase dispersa (Li *et al.*, 2016).

2.3.1 Correlações gerais de viscosidade de dispersões diluídas

Einstein (1911) foi pioneiro ao propor uma correlação para a viscosidade relativa, aplicável a sistemas de suspensão diluída. Sua equação revela que existe uma relação linear positiva entre a viscosidade relativa de emulsões e a fração volumétrica da fase dispersa, como mostra a Equação (3).

$$\mu_r = 1 + 2{,}5\varphi \tag{3}$$

A abordagem da Equação (3) retorna bons resultados quando aplicada para baixos valores da fração volumétrica, $\varphi < 0{,}01$.

Outra equação para suspensões concentradas é a equação de Brinkman (1952), algumas vezes também atribuída a Roscoe (1952).

$$\mu_r = (1 - \varphi)^{-2,5} \tag{4}$$

Outros modelos para cálculo da estimativa da viscosidade foram desenvolvidos. Na literatura, a maioria das demais equações de viscosidade consiste em expansões de virial de natureza empírica, tendo o formato de (5) (SCHRAMM, 2014).

$$\mu_r = 1 + a_0\varphi + a_1\varphi^2 + a_2\varphi^3 + \cdots \tag{5}$$

Para equações desse tipo, a_i são constantes empíricas. A exemplo, Guth–Gold–Simha desenvolveram a Equação (6) para dispersões com φ < 0.06 e Oliver–Ward desenvolveram a Equação (7) para esferas.

$$\mu_r = 1 + 2,5\varphi + 14,1\varphi^2 \tag{6}$$

$$\mu_r = 1 + a\varphi + a^2\varphi^2 + a^3\varphi^3 + \cdots \tag{7}$$

onde a é uma constante empírica.

2.3.2 Outros modelos desenvolvidos para emulsões

Algumas equações foram criadas com foco exclusivo em emulsões, incluindo variações da fórmula original de Einstein, como a proposta por Taylor (1932):

$$\mu_r = 1 + 2,5\varphi \left[\frac{\left(\left[\frac{\mu_2}{\mu_1} \right] + 0,4 \right)}{\left(\left[\frac{\mu_2}{\mu_1} \right] + 1 \right)} \right] \tag{8}$$

onde μ_2/μ_1 é a razão entre as viscosidades da fase dispersa (μ_2) e da fase contínua (μ_1).

Schramm (2014) comenta que outros pontos de partida frequentemente usadas são a equação de Richardson (1933) ou a de Broughton-Squires (1938), dadas pela Equação (9) e pela Equação (10), respectivamente.

$$\mu_r = e^{a\varphi} \tag{9}$$

$$\mu_r = e^{(a_1\varphi + a_2)} \tag{10}$$

sendo $a\ e\ a_i$ constantes empíricas.

Para emulsões concentradas (ϕ > 0,5), há a equação de Hatschek (1911):

$$\mu_r = \frac{1}{(1-\varphi^{1/3})} \tag{11}$$

e, em uma adaptação subsequente, surge a equação de Sibree (1931):

$$\mu_r = \left(1 + \left[\frac{12{,}5\varphi}{(1-a\varphi)}\right]\right)^2 \tag{12}$$

No caso de emulsões polidispersas, com possível comportamento não newtoniano, Eilers (1943) elaborou:

$$\mu_r = \left(1 + \left[\frac{12{,}5\varphi}{(1-a\varphi)}\right]\right)^2 \tag{13}$$

nesse contexto, o valor de a é 1,35 para esferas uniformes, enquanto, em outros casos, varia entre 1,28 e 1,30.

Pal & Rhodes (1989) desenvolveram modelos empíricos e teóricos para explicar o comportamento de viscosidade/concentração em emulsões newtonianas e não newtonianas com concentrações de fase dispersa abaixo de 74% em volume, sendo as equações propostas tanto aplicáveis para emulsões O/A quanto de A/O. Para o desenvolvimento da correlação, os autores utilizaram dois conjuntos de dados da literatura e 14 (quatorze) conjuntos de dados gerados por eles. A equação que apesentou o melhor ajuste aos dados é apresentada pela Equação (14).

$$\mu_r = \left(1 + \left[\frac{(\varphi/\varphi^*)}{(1{,}187 - (\varphi/\varphi^*))}\right]\right)^{2{,}49} \tag{14}$$

em que φ^* representa a fração volumétrica da fase dispersa na qual a viscosidade relativa chega a 100.

Diversas adaptações das equações de Richardson e Broughton–Squires surgiram para atender a necessidades específicas em diferentes campos. Um exemplo é a variação da equação

de Broughton–Squires, aplicada à viscosidade de emulsões de água em óleo cru, como o óleo cru do Mar do Norte. Nesse contexto, Rønningsen (1995) propôs uma correlação para viscosidade de emulsões do tipo A/O como função da fração volumétrica da fase dispersa e temperatura, utilizando oito tipos distintos de óleos brutos do Mar do Norte e ajustadas a uma equação empírica abrangente, na forma da Equação (15).

$$ln(\mu_r) = a_1 + a_2T + a_3\varphi + a_4T\varphi \quad (15)$$

Nessa expressão, a_i representam constantes empíricas dependentes da taxa de cisalhamento e T é a temperatura em °C. Neste trabalho, as viscosidades aparentes de emulsões de água em óleo bruto foram analisadas e a equação permitiu prever viscosidades relativas em faixas de temperatura de 5 a 40 °C, cortes de água variando entre 10 e 60%, e taxas de cisalhamento de 30 a 500 s^{-1}, com estimativas geralmente dentro de 1 a 30% dos valores obtidos experimentalmente (Rønningsen, 1995).

Com base em seis tipos distintos de óleos brutos, Farah *et al*. (2005) avaliou as viscosidades efetivas de várias emulsões sintéticas de água em óleo à pressão atmosférica, baseando-se na correlação para o cálculo da viscosidade cinemática D-341 utilizada pela *American Society for Testing and Materials* (*ASTM-2001*), originalmente usado para descrever a viscosidade cinemática em função da temperatura. A equação tomada como base foi ampliado para incorporar a variação da fração volumétrica da fase dispersa. Os testes foram conduzidos com temperaturas entre 8 e 50°C, fração volumétrica da fase dispersa entre 10 e 40% e taxas de cisalhamento entre 10 e 80s^{-1}, com óleos que tinha o °API variando entre 15,7 e 40,9.

$$ln\big(ln(v_e + 0{,}7)\big) = k_1 + k_2\varphi + k_3ln(T) + k_4\varphi ln(T) \quad (16)$$

Na Equação (16), v_e representa a viscosidade cinemática da emulsão (razão entre a sua viscosidade e densidade), k_i são constantes empíricas que dependem da taxa de cisalhamento e devem ser ajustados para cada emulsão e T é a temperatura em °C. A equação ajustada demonstrou ótima compatibilidade com os valores experimentais das viscosidades de emulsões de água em óleo, considerando temperatura e fração volumétrica de água.

2.4 INTELIGÊNCIA ARTIFICIAL E APRENDIZADO DE MÁQUINA

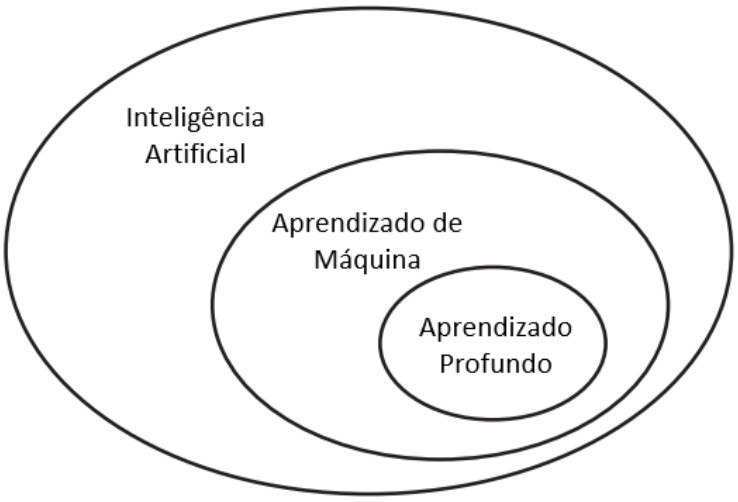
De acordo com Theodoridis (2020), os termos inteligência artificiai e aprendizado de máquina estão sendo cada vez mais utilizados para descrever o tipo de tecnologia de automação empregada na produção industrial, na distribuição de bens no comércio, no setor de serviços e em transações econômicas.

2.4.1 Inteligência Artificial

Conforme Belyadi e Haghighat (2021), a inteligência artificial é uma área da ciência da computação dedicada a estudar e desenvolver a capacidade dos computadores de simular processos cognitivos humanos, como aprendizado, raciocínio lógico, resolução de problemas e autocorreção. Em termos simples, trata-se da aplicação de inteligência computacional em substituição à inteligência humana, com o objetivo principal de criar sistemas inteligentes capazes de operar, reagir e replicar funções cognitivas humanas.

Por ser uma área abrangente, abordagens surgiram dentro do campo da inteligência artificial, entre elas está o aprendizado de máquina, com subáreas específicas como a aprendizado profundo (Belyadi; Haghighat, 2021; Chollet, 2021). A Figura 8 ilustra a relação entre inteligência artificial, aprendizado de máquina e aprendizado profundo.

Figura 8 - Inteligência artificial, aprendizado de máquina e aprendizado profundo.



Fonte: Adaptação de Chollet (2018).

Esses conceitos se organizam em camadas: a Inteligência Artificial é a área maior, que busca criar máquinas capazes de pensar e agir como humanos. Dentro dela, o Aprendizado de

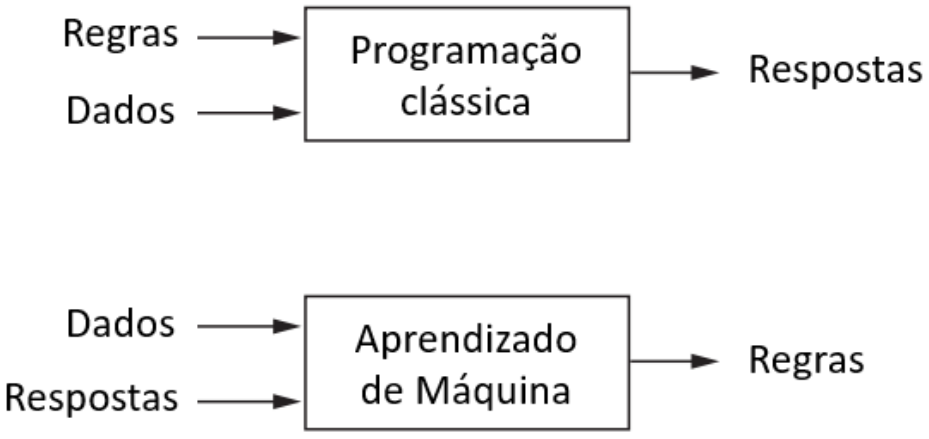
Máquina é um método onde os computadores aprendem com informações, reconhecendo padrões sem receber instruções diretas. Por fim, o Aprendizado Profundo é uma técnica específica do Aprendizado de Máquina, que usa estruturas complexas de redes neurais para lidar com grandes volumes de dados e resolver problemas sofisticados, como o reconhecimento de imagens e fala, sendo uma das ferramentas que impulsionam o progresso da inteligência artificial atualmente.

2.4.2 Aprendizado de Máquina

O aprendizado de máquina trata-se de uma vasta gama de métodos distintos que tem como objetivo criar modelos com base exclusivamente em dados empíricos. Esses métodos não necessitam da aplicação de leis físicas, nem da especificação de características da máquina. Eles identificam a dependência entre variáveis utilizando apenas os dados disponíveis (Bangert, 2021).

Chollet (2018) destaca que na programação clássica, os humanos definem regras explícitas (o programa) e fornecem dados para serem processados de acordo com essas regras. O resultado é um conjunto de respostas geradas estritamente com base nas instruções fornecidas. Essa abordagem depende fortemente de lógica predefinida e resolução de problemas estruturada, em contraste com os paradigmas modernos de IA, como o aprendizado de máquina, onde os sistemas aprendem padrões a partir dos dados, em vez de depender apenas de regras programadas. A Figura 9 resume a abordagem da programação clássica e do aprendizado de máquina.

Figura 9 - Paradigmas da programação clássica e aprendizado de máquina



Fonte: Adaptação de Chollet (2018).

Portanto, a questão central no aprendizado de máquina e no aprendizado profundo é transformar os dados de maneira significativa. Em outras palavras, é sobre aprender representações úteis dos dados de entrada que nos aproximem da saída esperada (Bangert, 2021).

Segundo Mohri *et al*. (2018), as tarefas em que o aprendizado de máquina tem sido amplamente estudado para aplicação são:

- **Classificação**: Trata-se de atribuir uma categoria a cada item. Por exemplo, a classificação de documentos, de imagens, de texto ou de reconhecimento de fala.

- **Regressão**: Refere-se à previsão de um valor real para cada item. Exemplos incluem a previsão de valores de ações ou de variações em indicadores econômicos. Na regressão, a penalidade por uma previsão incorreta depende da magnitude da diferença entre o valor real e o previsto, diferentemente da classificação, na qual não existe proximidade entre as categorias.

- **Ranqueamento**: Consiste em aprender a ordenar itens com base em algum critério. A busca na web, como a apresentação de páginas relevantes a uma consulta. Problemas semelhantes surgem no design de sistemas de extração de informações ou de processamento de linguagem natural.

- **Agrupamento (*Clustering*)**: Trata-se em dividir um conjunto de itens em subconjuntos homogêneos. Essa técnica é comumente usada para analisar grandes conjuntos de dados. Por exemplo, algoritmos de agrupamento identificam comunidades naturais dentro de grandes grupos de pessoas.

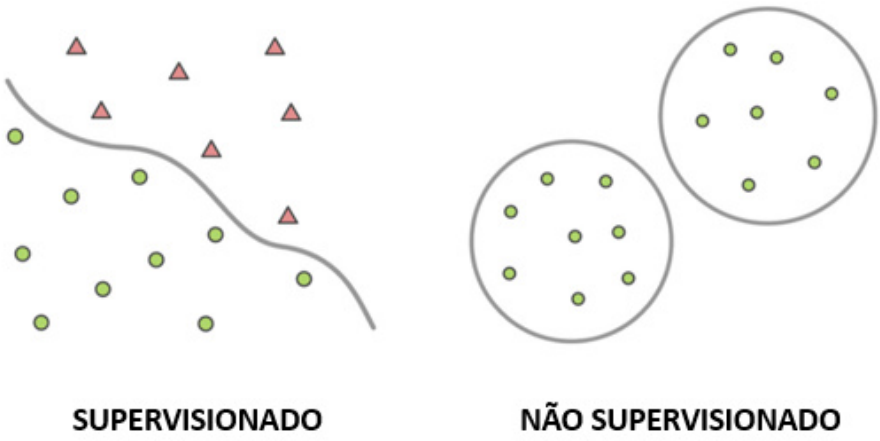
- **Redução de dimensionalidade ou aprendizado de representações (*Manifold Learning*)**: Envolve transformar a representação inicial dos itens em uma versão de menor dimensão, mantendo propriedades essenciais da representação original. Um exemplo comum é o pré-processamento de imagens digitais em tarefas de visão computacional.

2.4.3 Métodos de aprendizado de máquina

Os métodos de aprendizado de máquina podem ser organizados em dois critérios principais. O primeiro os classifica como supervisionados ou não supervisionados, enquanto o segundo os distingue entre métodos de classificação e de regressão (Bangert, 2021; Fan *et al.*, 2025).

Os métodos supervisionados utilizam conjuntos de dados nos quais se dispõe de informações empíricas tanto sobre as entradas do modelo quanto sobre os resultados esperados. Já os métodos não supervisionados operam com conjuntos de dados que contêm apenas as entradas, sem informações sobre os resultados desejados (Bangert, 2021; Theodoridis, 2020). A Figura 10 ilustra as diferenças entre os métodos supervisionados e não supervisionados.

Figura 10 - Ilustração da diferença entre aprendizado supervisionado e não supervisionado



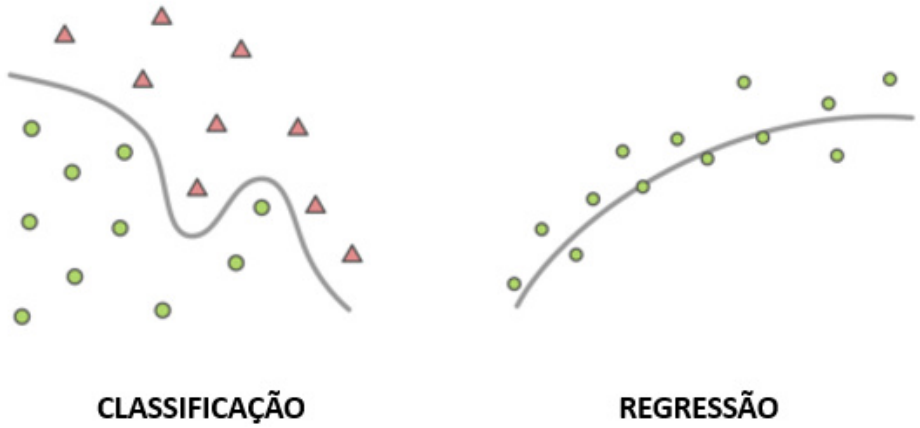
Fonte: Adaptação de Bangert (2021).

No aprendizado supervisionado, sabe-se previamente que existem dois grupos (círculos e triângulos) e é ensinado ao modelo a traçar uma divisão clara entre eles com base nos dados rotulados. No aprendizado não supervisionado, não há categorias pré-definidas, devendo o modelo analisar os pontos e descobrir padrões, agrupando-os em dois clusters de forma que os pontos dentro de cada cluster sejam muito semelhantes, e os pontos entre clusters sejam bastante diferentes, de acordo com uma métrica de similaridade relevante.

A classificação e a regressão são duas tarefas distintas no aprendizado supervisionado, diferenciadas pela natureza das variáveis que tratam. Na classificação, o objetivo é separar os dados em categorias ou classes distintas, associadas a variáveis categóricas, como no caso de distinguir círculos de triângulos. Já a regressão foca na modelagem de relações entre variáveis

de entrada e um valor numérico contínuo, como prever uma medida exata com base nos dados fornecidos (Bangert, 2021). A Figura 11 ilustra essa distinção de forma visual.

Figura 11 - Ilustração das diferenças entre métodos de classificação e regressão.



Fonte: Adaptação de Bangert (2021).

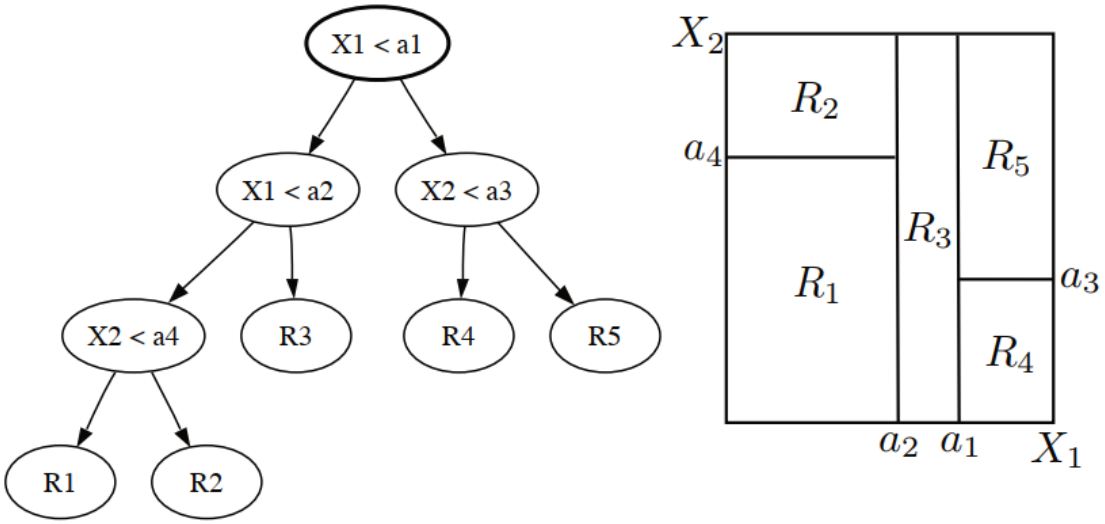
Essas abordagens são escolhidas com base na natureza do problema em questão e na estrutura dos dados disponíveis. A classificação lida com decisões discretas, enquanto a regressão trabalha com estimativas contínuas.

2.4.4 Modelos de aprendizado de máquina

2.4.4.1 Árvores de Decisão

A árvore de decisão é um algoritmo de aprendizado de máquina supervisionado que pode ser aplicado tanto em problemas de classificação quanto de regressão. Esse método divide os dados em subárvores, que, por sua vez, são subdivididas em outras subárvores (Belyadi; Haghighat, 2021). A Figura 12 mostra uma árvore de decisão binária em que é apresentado um exemplo simples no caso de um espaço bidimensional baseado em duas características, X1 e X2, assim como a partição que ele representa.

Figura 12 - Representação de uma árvore de decisão e a partição do espaço bidimensional induzido por essa árvore de decisão.



Fonte: Adaptação de Mehryar *et al*. (2018).

O nó raiz está localizado no nível mais alto e representa toda a população de dados. Já os nós de decisão, também chamados de nós internos, possuem duas ou mais ramificações. Por fim os nós terminais, conhecidos como nós folha, estão no nível mais baixo e não se dividem mais. É importante destacar que o termo "divisão" se refere ao processo de separar um nó em dois ou mais subnós (Belyadi; Haghighat, 2021).

Em um conjunto de dados com N atributos, determinar quais atributos devem ocupar o nó raiz ou os nós internos pode ser uma tarefa complexa e desafiadora. Belyadi e Haghighat (2021) descrevem alguns dos critérios mais relevantes para a seleção de atributos, sendo:

1. Entropia: trata-se de uma medida de incerteza ou pureza, calculada da seguinte forma:

$$E(S) = \sum_{i=1}^{c} -P_i log_2 P_i \tag{17}$$

em que "c" representa o número de classes e "Pi" é a probabilidade de uma classe dentro do conjunto de dados. Para múltiplos atributos, a entropia pode ser determinada da seguinte maneira:

$$E(X,Y) = \sum_{c \in Y} -P(c)E(c) \tag{18}$$

onde "X" representa o estado atual e "Y" é o atributo selecionado. P(c) corresponde à probabilidade do atributo, enquanto E(c) refere-se à entropia do atributo.

2. Ganho de informação (GI): O ganho de informação indica o quão eficazmente um atributo consegue separar os exemplos de treinamento de acordo com sua classificação alvo.

$$GI(X,Y) = E(Y) - E(Y,X) \tag{19}$$

ao construir uma árvore de decisão, é essencial identificar um atributo que proporcione o maior ganho de informação e a menor entropia.

3. Índice Gini: O índice Gini mede o grau de heterogeneidade dos dados em um nó. Ele avalia o quão "misturadas" estão as classes no conjunto de dados. Quanto maior o índice Gini, maior a heterogeneidade (ou mistura) dos dados. Por outro lado, um índice Gini baixo indica maior homogeneidade, ou seja, os dados estão mais concentrados em uma única classe.

$$Índice\ Gini = 1 - \sum_{i=1}^{c} (P_i)^2 \tag{20}$$

onde "P_i" representa a probabilidade de um elemento ser classificado em uma determinada classe. Em exemplos perfeitamente classificados, o índice de Gini seria igual a 0.

Um dos maiores desafios ao utilizar uma árvore de decisão é o sobreajuste. Uma das estratégias para evita-la é a poda.

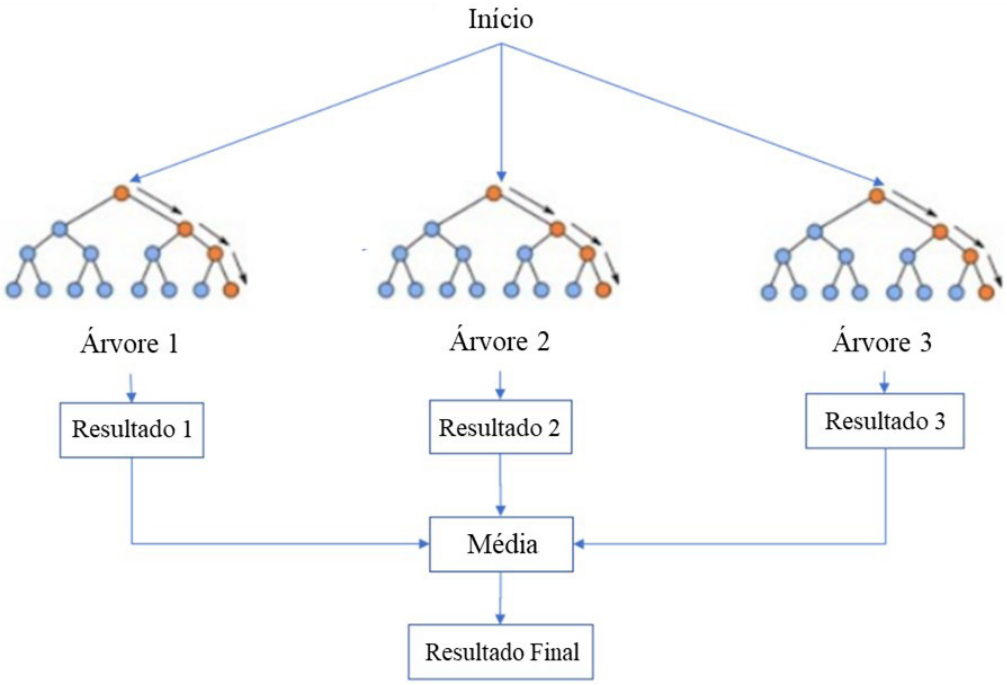
4. Poda: consiste em remover ramos ou partes da árvore que não contribuem significativamente para a classificação das instâncias ou que comprometem a precisão geral. Além disso, é fundamental empregar validação cruzada ao trabalhar com árvores de decisão, garantindo que o modelo não esteja excessivamente ajustado aos dados.

2.4.4.2 Floresta Aleatória

A Floresta Aleatória (Random Forest, ou RF) é uma técnica de aprendizado conjunto que constrói diversas árvores de decisão durante o treinamento e produz uma previsão média com base nas estimativas de cada árvore individual (Pandey *et al.*, 2020; Wang; Chakraborty; Chakraborty, 2021).

Esse algoritmo, fundamentado em um conjunto (*ensemble*) de árvores de decisão, utiliza o método de *bagging*, aplicando-o tanto às características de entrada quanto aos dados de treinamento. Assim, para cada árvore de decisão gerada no modelo de floresta aleatória, um subconjunto de atributos de entrada e dados de treinamento é selecionado aleatoriamente para seu treinamento, com várias árvores sendo treinadas em paralelo de forma independente (Pandey *et al.*, 2020). A estrutura básica desse método é ilustrada na Figura 13.

Figura 13 - Estrutura simplificada do modelo Floresta Aleatória.



Fonte: Adaptação Wang *et al*. (2021).

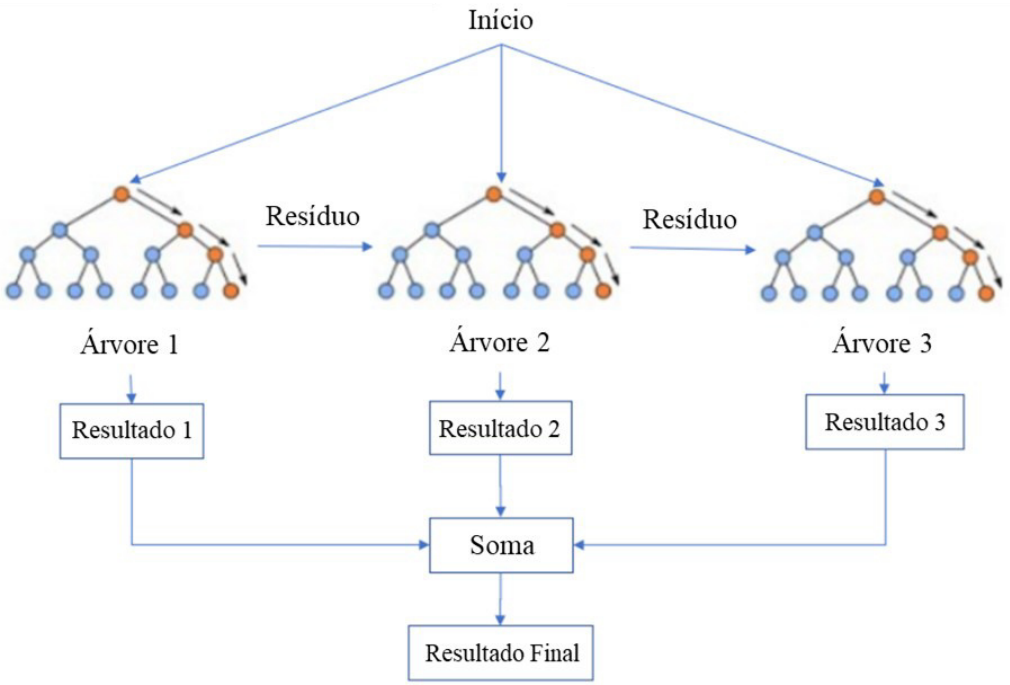
Ao final, os resultados de todos os submodelos são combinados por meio de uma média. Além de utilizar subconjuntos distintos dos dados para construir cada árvore, as florestas aleatórias diferenciam-se pelo modo como as árvores são formadas. Em árvores de decisão tradicionais, os nós são divididos com base na melhor escolha entre todas as variáveis, buscando reduzir a entropia resultante da separação dos dados representados pelo nó pai. Já na floresta

aleatória, os pontos de divisão em cada nó são selecionados aleatoriamente entre as melhores opções dentro de um subgrupo de preditores. Esse processo ajuda a prevenir o sobreajuste, um problema frequente em árvores de decisão únicas e muito profundas (Belyadi; Haghighat, 2021; Pandey *et al.*, 2020; Wang; Chakraborty; Chakraborty, 2021).

2.4.4.3 eXtreme Gradient Boosting

O *eXtreme Gradient Boosting* (XGBoost) é um sistema de aprendizado de máquina focado em *tree boosting*, amplamente reconhecido por seu impacto em desafios de aprendizado de máquina e mineração de dados. Diferentemente do Random Forest, o XGBoost constrói árvores de decisão sequencialmente, com cada modelo ajustando os erros residuais deixados pelos anteriores (Chen; Guestrin, 2016; Wang; Chakraborty; Chakraborty, 2021). A estrutura básica deste método é ilustrada na Figura 14.

Figura 14 - Estrutura simplificada do XGBoost



Fonte: Adaptação de Wang *et al*. (2021).

Sua eficiência e robustez destacam-se especialmente em dados estruturados, sendo amplamente utilizado em problemas de classificação e regressão. Com sua capacidade de entregar resultados, o XGBoost tornou-se popular em competições como *Kaggle*. Ele oferece recursos otimizados para lidar com diversos de exemplos e é reconhecido por sua habilidade de

superar desafios complexos, consolidando-se como uma ferramenta indispensável na ciência de dados e aprendizado de máquina (Chen; Guestrin, 2016).

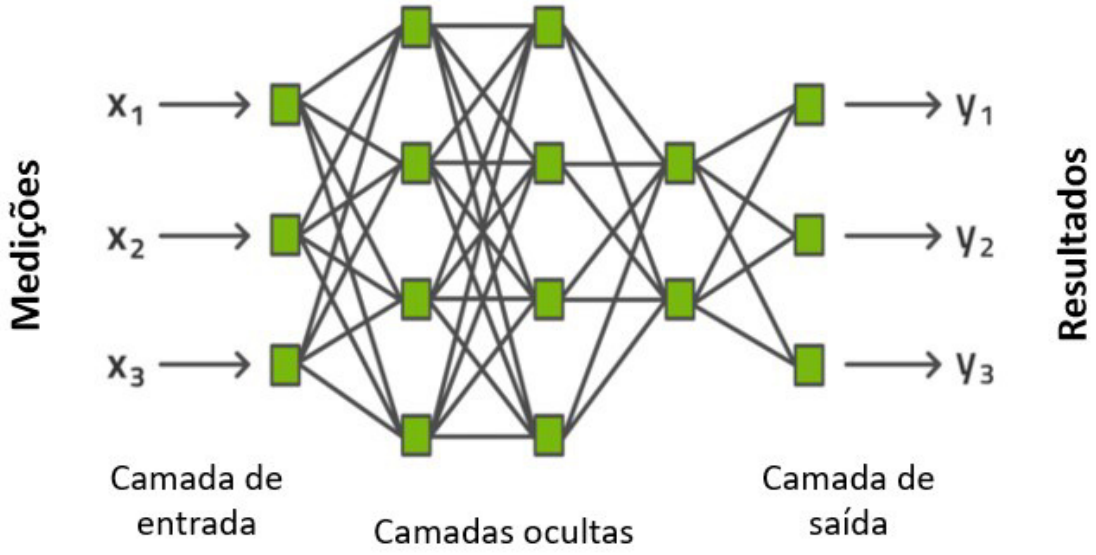
No aprendizado de máquina, o principal fator para o sucesso de um modelo é a função objetivo utilizada no processo de aprendizado. Essa função é composta por dois elementos: a função de perda e a regularização. A função de perda mede a diferença entre os valores previstos pelo modelo e os valores reais (por exemplo, o erro quadrático médio em um problema de regressão). Enquanto isso, a regularização gerencia a complexidade do modelo, ajudando a prevenir o problema de sobreajuste (*overfitting*). Esse mecanismo é impulsionado ao utilizar o gradiente descendente para minimizar a função de perda, incluindo regularização por termos L1 (*Lasso*), que tende a zerar pesos de características menos importantes, e L2 (*Ridge*), que penaliza pesos grandes. Ambos melhoram a generalização do modelo e evitam o *overfitting* (Ahmetoglu; Das, 2022; Chen; Guestrin, 2016; Wang; Chakraborty; Chakraborty, 2021).

2.4.4.4 Redes Neurais Artificiais

Segundo Bangert (2021), as redes neurais é talvez a técnica mais famosa e mais utilizada no arsenal de aprendizado de máquina. Com o avanço recente do aprendizado profundo, esse modelo ganhou novas dimensões, graças a técnicas inovadoras para o treinamento de seus parâmetros.

As redes neurais artificiais tiveram sua origem inspirada no funcionamento de neurônios biológicos no cérebro. O processamento de informações nesses sistemas reflete os mecanismos dos neurônios presentes em organismos biológicos. Assim, os neurônios são os principais elementos que compõem redes neurais artificiais, passando sinais entre si, como ocorre no cérebro humano. Neurônios artificiais conectam múltiplos *inputs* (entradas) a uma ou várias saídas, utilizando pesos associados e funções de ativação não lineares. Durante o treinamento, dados são fornecidos às redes neurais, que, por meio de algoritmos específicos, determinam os pesos ideais para caracterizar o comportamento das saídas em relação aos diversos *inputs* (Belyadi; Haghighat, 2021).

Figura 15 - Modelo esquemático de uma Rede Neural Artificial.



Fonte: Adaptação de Bargert (2021).

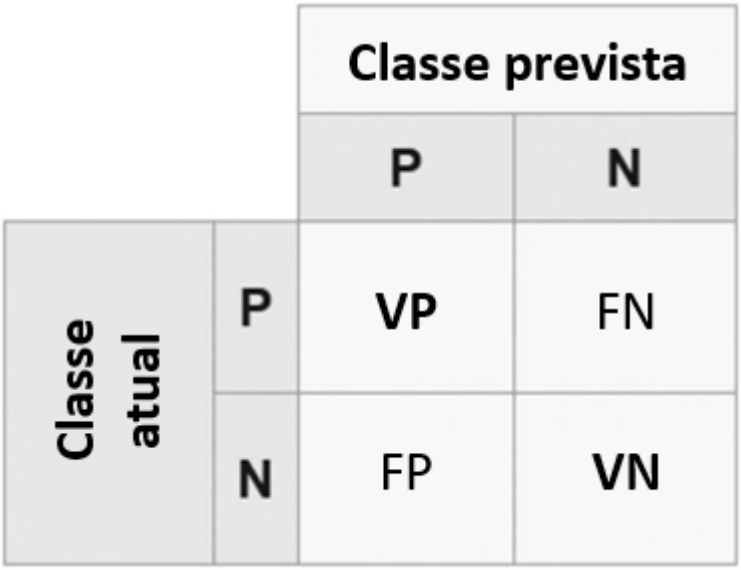
As redes neurais oferecem um alto potencial para explorar e analisar grandes bancos de dados históricos que não parecem ser utilizados em modelagem convencional. As redes neurais são organizadas em uma sequência de neurônios dispostos em camadas. Cada nó, semelhante a um neurônio biológico, é atribuído a variáveis de entrada e saída. Esses nós formam diferentes camadas: camada de entrada, uma ou mais camadas ocultas e camada de saída, conectando-se entre si, como ilustrado na Figura 15. A estrutura dessas redes, conhecida como topologia, determina tanto a quantidade de camadas ocultas quanto o número de nós em cada camada, necessários para vincular a entrada à camada de saída.

2.4.5 Matriz de Confusão ou Matriz de Erro

Trata-se de uma matriz $N \times N$, onde a diagonal principal representa as N classificações corretas e as entradas fora da diagonal correspondem aos erros potenciais. Esse método oferece uma forma visual de analisar a eficácia de um modelo de aprendizado de máquina em tarefas de classificação (Pandey *et al.*, 2020).

A Figura 16 apresenta uma ilustração de matriz de confusão aplicada à classificação binária, que abrange duas classes.

Figura 16 - Matriz de confusão clássica para um problema de classificação binário



Fonte: Adaptação de Pandey (2020).

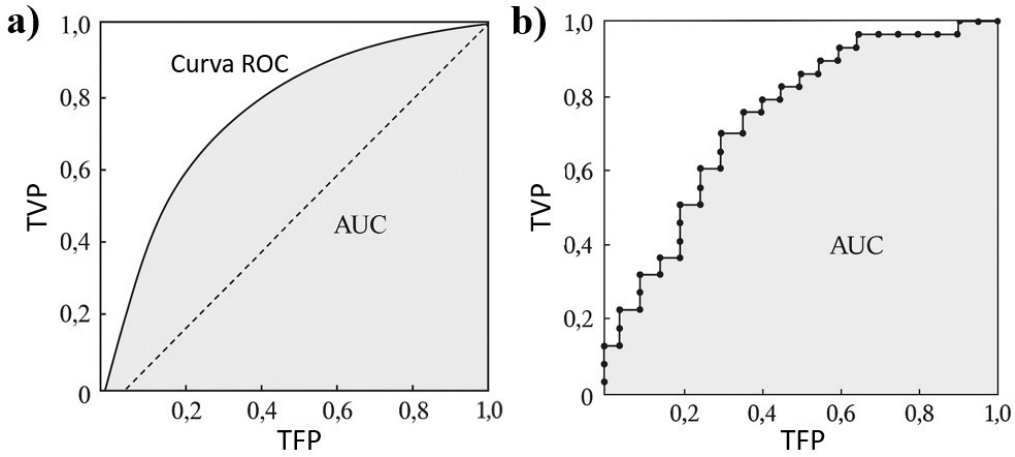
onde P = positivo, N = negativo, VP = verdadeiro positivo; FP = falso positivo; VN = verdadeiro negativo; FN = falso negativo.

2.4.6 Curvas Características de Operação do Receptor

Uma curva Característica de Operação do Receptor (*Receiver Operating Characteristic - ROC*) é um gráfico utilizado para demonstrar a desempenho de um sistema de classificação binária em diferentes valores de corte (Arjaria; Rathore; Cherian, 2021).

É um método amplamente utilizado para avaliar, comparar e selecionar classificadores, oferecendo uma forma eficaz de visualizar seu desempenho preditivo, auxiliando na seleção de um ponto de operação ou limiar de decisão adequado. Sua primeira aplicação conhecida foi durante a Segunda Guerra Mundial, quando foi empregada no processamento de sinais de radar, onde operadores de radar avaliavam se os sinais na tela eram objetos reais ou apenas ruídos. Posteriormente, esse método foi incorporado à teoria de detecção de sinais para ilustrar o equilíbrio entre taxas de acerto e taxas de falso alarme de classificadores (Bradley, 1997; Fan; Upadhye; Worster, 2006; Majnik; Bosnić, 2013; Zhou, 2021). A Figura 17 ilustra como é a apresentação de uma curva ROC.

Figura 17 - Curvas ROC.



Fonte: Adaptado de Zhou (2021).

A curva ROC é construída com base na relação entre a taxa de verdadeiros positivos (TVP) e a taxa de falsos positivos (TFP), permitindo a análise comparativa de diferentes classificadores (Arjaria; Rathore; Cherian, 2021). As Equações (21) e (23) mostra como é o cálculo dessas taxas.

$$TVP = Sensibilidade = \frac{VP}{VP + FN} \tag{21}$$

$$Especificidade = \frac{VN}{VN + FP} \tag{22}$$

$$TFP = 1 - Especificidade = \frac{FP}{FP + VN} \tag{23}$$

A escolha do valor de corte ideal em um teste classificatório quase sempre envolve um equilíbrio entre sensibilidade (capacidade de identificar verdadeiros positivos) e especificidade (capacidade de evitar falsos positivos). Como essas métricas variam com cada valor de corte, visualizar o ponto ideal pode ser desafiador. A curva ROC fornece uma representação gráfica que facilita a compreensão desse equilíbrio em qualquer processo de classificação baseado em variáveis contínuas, permitindo identificar como diferentes valores de corte afetam o desempenho geral do classificador e ajudando na escolha mais apropriada para o contexto específico (Fan; Upadhye; Worster, 2006).

Fan *et al*. (2006) ainda comenta que em um cenário ideal, o melhor valor de corte é aquele que maximiza tanto a sensibilidade quanto a especificidade, identificável na curva ROC

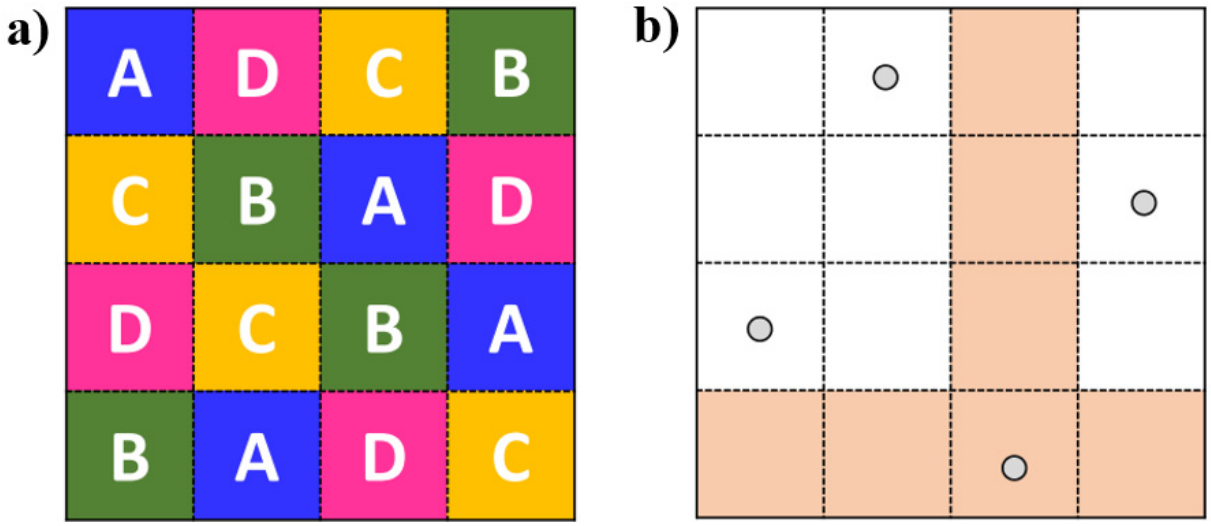
pelo ponto mais alto no eixo vertical e mais à esquerda no eixo horizontal, localizado no canto superior esquerdo. No entanto, alcançar esse equilíbrio perfeito na prática é raro.

2.5 AMOSTRAGEM POR HIPERCUBO LATINO

O método de amostragem por hipercubo latino (*Latin Hypercube Sampling - LHS*), introduzido por McKay *et al*. (1979), é uma técnica estatística amplamente aplicada em simulações Monte Carlo e experimentos computacionais. Ele gera pontos de amostra distribuídos de forma mais uniforme em espaços multidimensionais, comparado à amostragem aleatória, garantindo estratificação cuidadosa na distribuição de probabilidade dos parâmetros de entrada. Isso permite maior eficiência na cobertura do espaço de parâmetros, resultados mais representativos e precisão nas análises (Navid; Khalilarya; Abbasi, 2018). Além disso, o LHS pode ser usado de forma independente para processos de otimização em planejamentos experimentais (*DOE – Design of Experiments*) (Ebbs-Picken; Da Silva; Amon, 2023).

Sua inspiração está nos quadrados latinos, um conceito da matemática combinatória em que uma matriz n x n é preenchida com n objetos distintos (como números ou símbolos), de forma que cada objeto ocorra apenas uma vez em cada linha e coluna. O termo “Latino” foi inspirando pelo trabalho do famoso matemático Leonhard Euler, que usava caracteres latinos como elementos nos quadrados. A técnica LHS adapta esse princípio para espaços bidimensionais e amostras de tamanho n. Cada dimensão é dividida em n intervalos uniformes com probabilidade marginal de $1/n$. A amostragem ocorre aleatoriamente dentro de cada intervalo, garantindo distribuição uniforme e evitando sobreposição de pontos em níveis idênticos (Sheikholeslami; Razavi, 2017). Um exemplo prático envolve a distribuição de 4 pontos em um espaço bidimensional, como mostra a Figura 18.

Figura 18 - Organização de pontos em um espaço bidimensional a partir da amostragem do hipercubo latino.



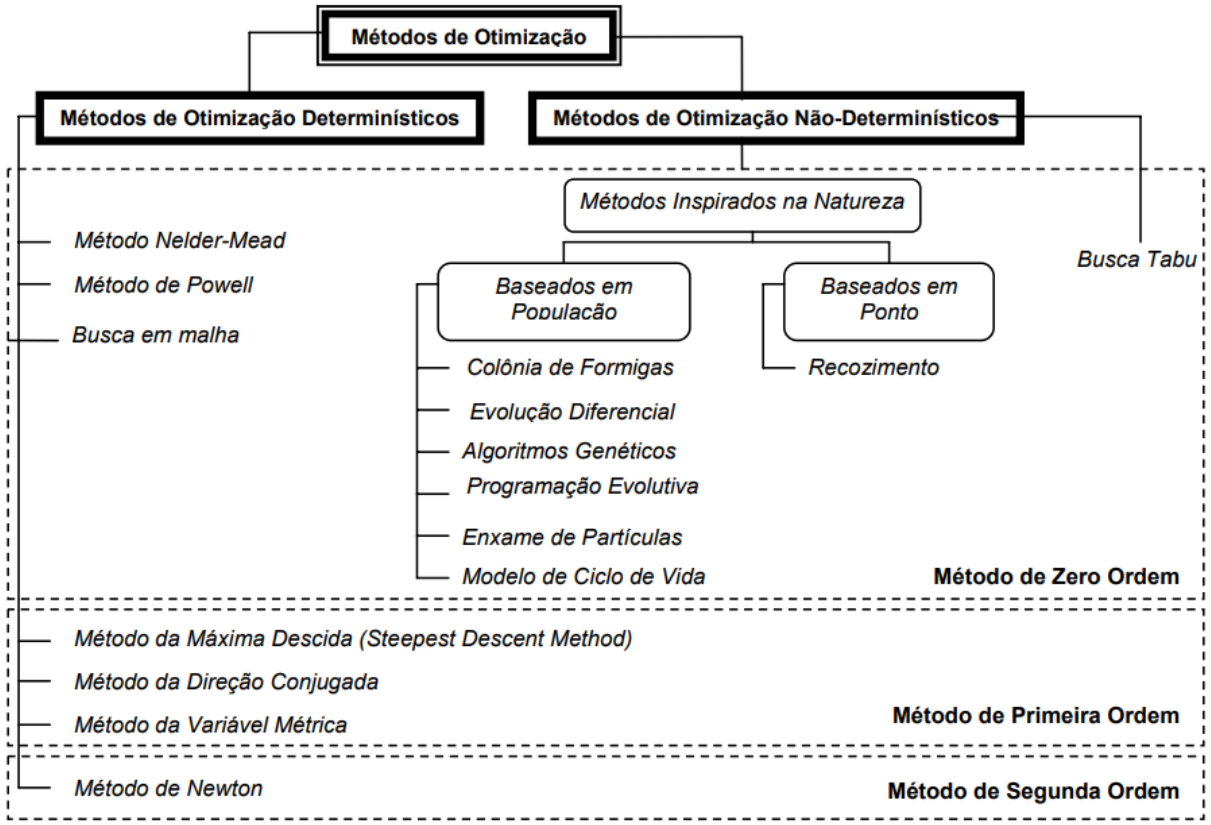
Fonte: Adaptado de Sheikholeslami *et al*. (2017).

O LHS garante que cada intervalo de cada dimensão tenha, no mínimo, um ponto amostral, promovendo uma cobertura abrangente do espaço de busca. Isso reduz o risco de regiões importantes ficarem sem representação (lacunas) ou de concentrar muitos pontos em uma mesma área (redundâncias).

2.6 MÉTODOS DE OTIMIZAÇÃO

A otimização de processos é essencial para aprimorar sistemas e operações, buscando alcançar o melhor desempenho possível, seja minimizando custos, maximizando eficiência ou melhorando a qualidade. Em muitas situações de problemas de aprendizado de máquina, as relações entre as entradas e saídas dos modelos possuem geralmente interações não lineares. Por conta disso, é possível que haja mais de uma solução ótima. Os métodos de otimização podem ser classificados em determinísticos e não determinísticos, cada um com características, vantagens e limitações distintas.

Figura 19 - Classificação dos métodos de otimização.



Fonte: Borges (2008).

Os métodos de otimização determinísticos ou clássicos se baseiam em informações de derivadas de primeira e segunda ordem para determinar a direção da busca, apresentando uma rápida convergência próxima ao ponto ótimo da função objetivo (Vanderplaats, 1999). Contudo, a ausência de continuidade nas funções a serem otimizadas ou nas restrições aplicadas, juntamente com a existência de funções não convexas e multimodais, pode gerar dificuldades numéricas no método, levando a estimativas que apontam para ótimos locais em vez de uma solução global.

Os métodos de otimização não determinísticos, estocásticos ou randômicos são algoritmos que se inspiram em fenômenos naturais, como a evolução de populações, processos físicos ou modelos matemáticos estruturais (Coelho; Krohling, 2003).

Figura 20 - Vantagens e desvantagens dos métodos de otimização não determinísticos.

Vantagens	Desvantagens
- Dispensam o uso de derivadas da função para direcionar a busca dos pontos ótimos; - Não investem todo esforço computacional em um único ponto, mas sim sobre uma população de pontos; - São reconhecidos como métodos de busca global, capazes de escapar de ótimos locais.	- Seus desempenhos variam de execução para execução, pelo fato de serem métodos estocásticos; - São muito mais demorados que os métodos clássicos do ponto de vista do número de avaliações da função objetivo.

Fonte: Pereira (2019).

2.6.1 Enxame de partículas (PSO)

O Enxame de Partículas (*Particle Swarm Optimization* - PSO) é um algoritmo de otimização estocástica baseado em população, proposta por Eberhart e Kennedy em 1995. É inspirado no comportamento coletivo inteligente de alguns animais, como insetos, manadas, pássaros e peixes (Wang; Tan; Liu, 2018). Esses enxames cooperam para encontrar alimento, e cada membro do grupo altera constantemente seu padrão de busca com base em suas próprias experiências de aprendizado e nas experiências de outros membros. A Figura 21 ilustra a analogia entre o comportamento de um enxame inteligente (representado por um bando de pássaros) e o modelo proposto, o qual busca otimizar a função de interesse ao simular a coreografia coletiva realizada pelas aves durante a busca por recursos.

Figura 21 - Identificação de termos do PSO.

Termo	Significado
Partícula	Pássaro
Enxame	Bando de pássaros
Espaço de busca	Área sobrevoada pelos pássaros
Posição	Localização de cada pássaro durante o voo
Solução ótima	Localização do pássaro onde ele encontrou o alimento
Fitness	Função de avaliação
pbest	Melhor posição conhecida pelo pássaro (experiência pessoal)
gbest	Melhor posição conhecida pelo enxame (experiência coletiva)

Fonte: Pereira (2019).

Segundo Lian *et al.* (2008), o PSO utiliza uma população chamada enxame, onde cada indivíduo dentro do enxame é denominado de partícula. Uma partícula i em uma iteração k se desloca através do espaço de busca com dois atributos:

- A posição atual dentro do espaço de busca N-dimensional $X_i^k = (x_1^k, \ldots, x_n^k, \ldots, x_N^k)$ do problema, com $x_n^{min} \leq x_n^k \leq x_n^{max}$ para todo $n \in [1, N]$, onde x_n^{min} e x_n^{max} são os limites da coordenada n.
- Sua velocidade que é representada vetorialmente por $V_i^k = (v_1^k, \ldots, v_n^k, \ldots, v_N^k)$ nesse mesmo espaço N-dimensional do problema.

Em cada iteração, a velocidade e a posição de todas as partículas são ajustadas com base nos dois melhores valores obtidos ao longo da busca. O primeiro, denominado *pbest*, representa o melhor valor já encontrado por uma partícula individualmente. O segundo, chamado *gbest*, corresponde ao melhor valor descoberto até o momento por qualquer membro da população no algoritmo PSO. Após determinar esses dois parâmetros, a atualização da posição e velocidade das partículas ocorre seguindo as Equações (24) e (29).

$$V_i^{k+1} = w \cdot V_i^k + c_1 \cdot r_1 \cdot \left(pbest_i^{k+1} - X_i^k\right) + c_2 \cdot r_2 \cdot \left(gbest^k - X_i^k\right) \tag{24}$$

$$X_i^{k+1} = X_i^k + V_i^{k+1} \tag{25}$$

em que r_1 e r_2 são números gerados aleatoriamente no intervalo entre [0, 1] e c_1 e c_2 são chamados de parâmetro cognitivo e social, respectivamente. O termo $c_1 \cdot r_1 \cdot \left(pbest_i^{k+1} - X_i^k\right)$ representa a distância entre a partícula (i) e sua melhor posição até a k-ésima iteração. Sua função é ampliar a exploração do espaço de busca, permitindo que o algoritmo investigue diferentes regiões em busca do mínimo global. Esse mecanismo evita que a solução fique presa em mínimos locais, promovendo uma busca mais eficiente e abrangente. Por outro lado, o termo $c_2 \cdot r_2 \cdot \left(gbest^{k+1} - X_i^k\right)$ representa a distância entre a partícula i e a melhor posição encontrada pela população até a k-ésima iteração.

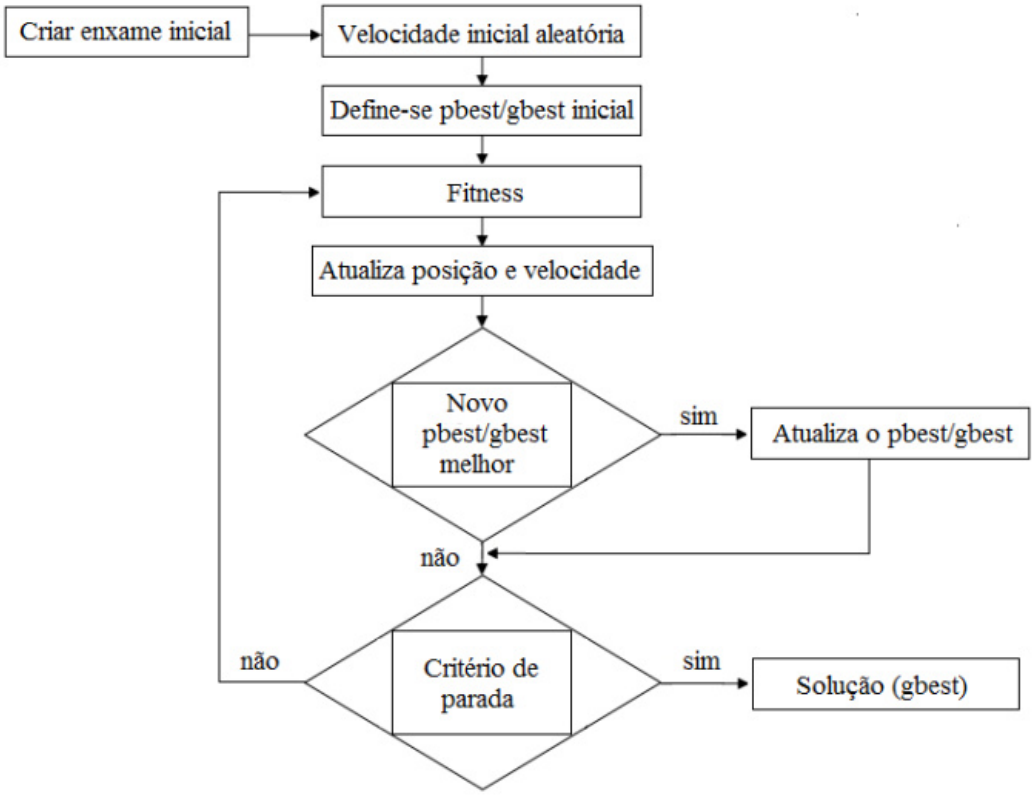
As Equações (26) e (27) definem como os melhores valores pessoais ($pbest_i^k$) e global ($gbest^k$) são no tempo k, respectivamente.

$$pbest_i^{k+1} = \begin{cases} pbest_i^k & se\ f\left(pbest_i^k\right) > f\left(x_i^{k+1}\right) \\ x_i^{k+1} & se\ f\left(pbest_i^k\right) \leq f\left(x_i^{k+1}\right) \end{cases} \quad (26)$$

$$gbest^{k+1} = min\left\{f\left(pbest_i^{k+1}\right), f(gbest^k)\right\} \quad (27)$$

Uma população inicial de partículas é proposta, com os vetores de velocidade e posição sendo gerados aleatoriamente. Em seguida, o cálculo do fitness é aplicado a cada partícula da população. Após essa avaliação, são determinados os valores de *pbest* (a melhor posição encontrada por cada partícula) e *gbest* (a melhor posição identificada pelo enxame). A partir das novas posições, se o critério de parada for atingido, a solução obtida para o problema é apresentada na Figura 22 (Pereira, 2019).

Figura 22 - Fluxograma do PSO.



Fonte: Pereira (2019).

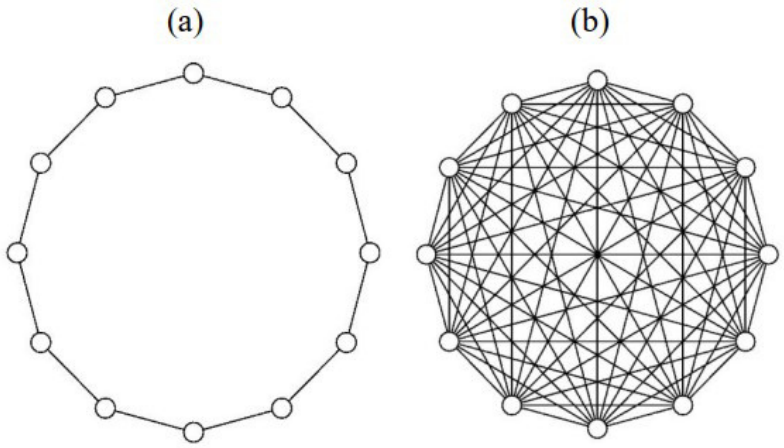
Assim, a evolução do algoritmo PSO está diretamente ligada à trajetória percorrida pelo enxame e ao tempo necessário para encontrar a melhor solução do problema. Nesse contexto, a escolha adequada do número de partículas e iterações, dos limites do espaço de

busca e das restrições de velocidade é essencial, pois cada parâmetro influencia o desempenho do algoritmo, trazendo vantagens e desafios conforme a natureza do problema (Pereira, 2019).

O número de partículas e de iterações influencia diretamente a probabilidade de o algoritmo encontrar a solução ótima para o problema. No entanto, um aumento nesses parâmetros implica em um maior número de testes e atualizações, o que, por sua vez, resulta em um tempo computacional mais elevado. Assim, é essencial equilibrar esses fatores para garantir eficiência sem comprometer a qualidade da solução (Andrade; Costa; Angélico, 2013).

O espaço de busca e a velocidade das partículas geralmente são restringidos por valores máximos e mínimos. Um aumento na velocidade permite que as partículas alcancem a solução ótima mais rapidamente, porém, pode limitar a exploração de regiões que contenham máximos ou mínimos locais importantes para a resolução do problema. Por outro lado, velocidades mais baixas aumentam o número de iterações necessárias para a convergência das partículas, proporcionando uma exploração mais detalhada do espaço de busca e elevando a probabilidade de encontrar uma solução global (Andrade; Costa; Angélico, 2013).

Figura 23 - Topologias: (a) local e (b) global.



Fonte: Rosendo (2010)

Na configuração de topologia local, as partículas de um enxame são dispostas em uma estrutura semelhante a um anel. Cada uma delas mantém dois vizinhos que mudam com o movimento do grupo, caracterizando uma vizinhança dinâmica. A comunicação ocorre somente entre vizinhos imediatos, o que torna o fluxo de informações mais lento. Por outro lado, quando se adota a topologia global, todas as partículas têm acesso à posição da melhor partícula identificada no espaço de busca, além de seu próprio histórico. Essa abordagem acelera o processo de convergência, embora não assegure que a solução encontrada seja necessariamente a mais precisa ou eficaz.

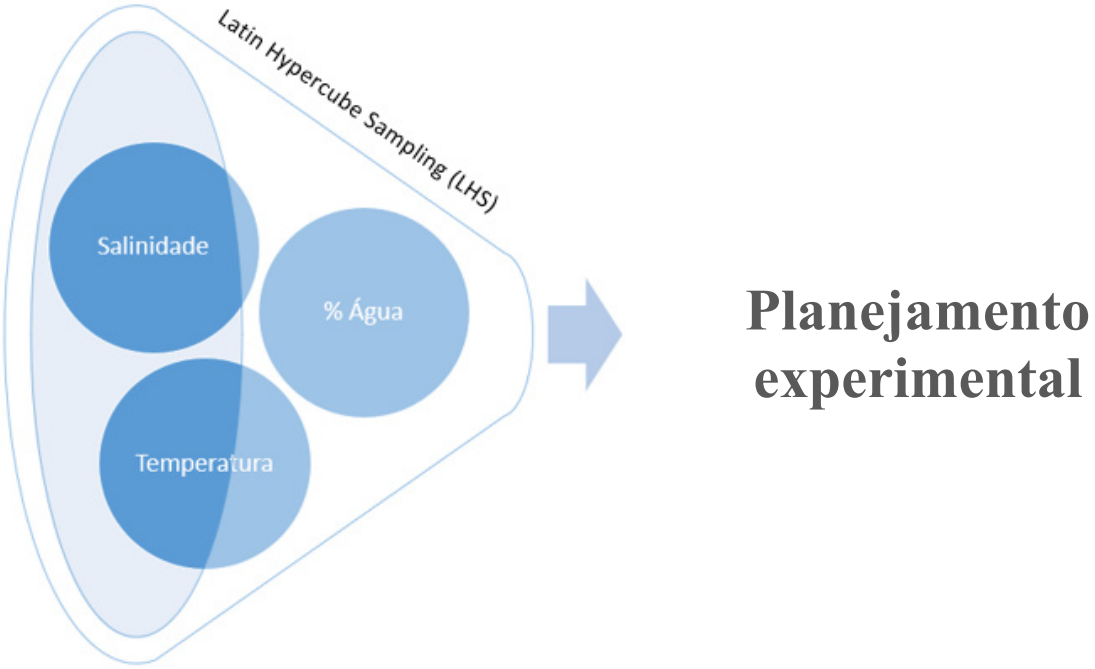
3 METODOLOGIA

Esta seção detalha a metodologia adotada para o desenvolvimento de modelos de aprendizado de máquina para predizer a estabilidade, viscosidade e tamanho de gotas de emulsões A/O. A metodologia partiu de um planejamento experimental, seguido pela obtenção dos dados experimentais através da caracterização das emulsões. Após a organização e o tratamento, esses dados foram utilizados para treinar os modelos em um processo que incluiu otimização de hiperparâmetros e validação cruzada. Por fim, os modelos foram avaliados por métricas estatísticas e suas predições interpretadas com análise da influência que os parâmetros que alimentaram os modelos tiveram sobre as predições.

3.1 PLANEJAMENTO EXPERIMENTAL

Visando otimizar o planejamento experimental e criar uma base de dados para alimentar modelos de Inteligência Artificial, aplicou-se o método LHS para gerar variados cenários e produzir valores de entrada para a estimativa das variáveis de saída.

Figura 24 - Esquema visual de geração do planejamento experimental.



Fonte: Próprio autor.

As variáveis de entrada analisadas foram temperatura (40-80°C), cortes de água (0-50%) e salinidade (40-240 g/L), intervalos definidos por serem condições típicas de separação das amostras de óleo utilizadas. Os intervalos definidos para essas variáveis foram empregados para gerar um total de 25 ensaios experimentais, que foram aplicados individualmente a cada um dos 13 óleos avaliados. Avaliaram-se a influência dessas condições na estabilidade, viscosidade aparente e o tamanho médio de gotas das emulsões.

3.2 AMOSTRAS DE ÓLEO

As amostras de óleo utilizadas neste trabalho foram caracterizadas quanto à gravidade API e à composição SARA, propriedades incluídas como variáveis de entrada nos modelos de aprendizado de máquina. Os principais resultados estão listados na Tabela 2.

Tabela 2 - Caracterização dos óleos utilizados no trabalho

Petróleo	**°API**	**Saturados ± 2, % (m/m)**	**Aromáticos ± 2, % (m/m)**	**Resinas ± 3, % (m/m)**	**Asfaltenos ± 0,05, % (m/m)**
P1	29,20	54,00	24,00	22,00	0,50
P2	29,50	53,10	25,60	21,13	0,17
P3	27,50	78,90	20,70	0,40	0,75
P4	29,00	52,10	30,10	17,80	0,50
P5	32,50	60,90	21,80	16,90	0,40
P6	24,40	60,70	18,90	19,10	1,30
P7	28,10	56,80	24,30	18,00	0,90
P8	28,60	62,90	18,40	17,90	0,70
P9	26,00	52,00	27,00	21,00	0,54
P10	19,00	40,00	20,00	38,00	2,09
P11	23,10	39,90	22,60	36,10	1,47
P12	20,40	42,40	24,50	30,00	3,16
P13	16,50	32,00	27,00	35,00	6,34

Fonte: Próprio autor.

As amostras de óleo apresentam uma variedade de características em relação à gravidade API. Dentre as 13 amostras avaliadas, 1 (uma) foi classificada como óleo leve (°API > 31), 9 (nove) como óleos médios (22 ≤ °API ≤ 31) e 3 (três) como óleos pesados (°API < 22). Essa diversidade é importante para garantir maior abrangência para a construção de modelos de IA com boa capacidade de generalização.

3.3 PREPARAÇÃO DE EMULSÕES

Emulsões do tipo A/O foram produzidas em diferentes temperaturas, salinidades e cortes de água de acordo com o planejamento experimental definidos pelo método LHS, mostrado na Tabela A.1 em anexo. A salmoura utilizada na formação de emulsões foi sintetizada a partir de água deionizada (condutividade de 18,2 ± 0,2 mΩcm, a 298,15 K) e cloreto de sódio (NaCl, Sigma-Aldrich, São Paulo, Brasil), em diferentes concentrações.

Para determinação da melhor combinação de velocidade e tempo de agitação para formação de emulsões, ensaios preliminares foram realizados tomado como base a metodologia usada por Feitosa (2018). Para isso, foram testadas combinações de diferentes velocidades (3000; 3400; 3600; 4000; 5000; 7000 rpm) e tempos de agitação (5, 10 e 15 minutos) na produção de 50 mL de emulsão, contendo 30% (v/v) de solução salina.

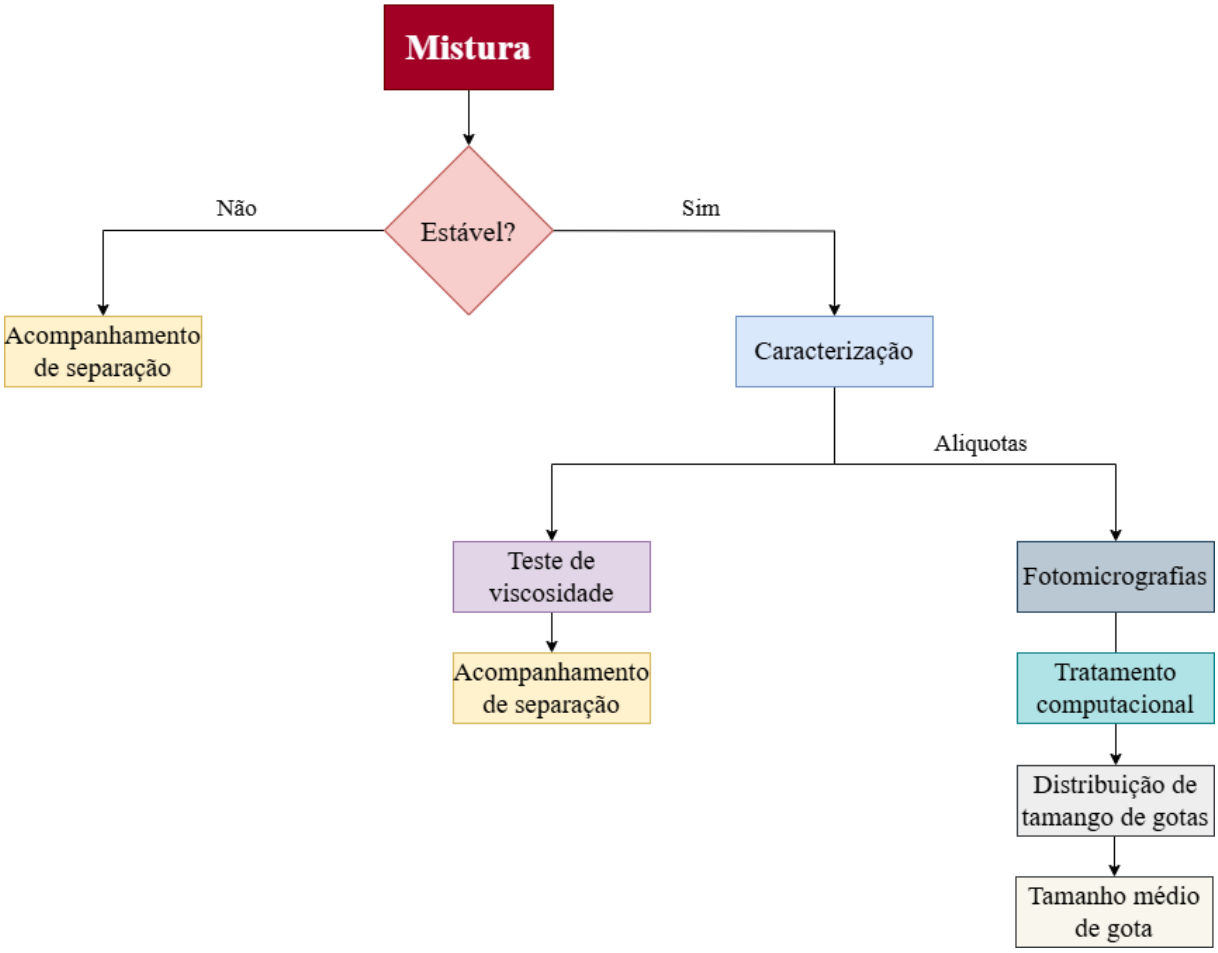
A menor velocidade possível combinada com o menor tempo de agitação que propiciasse uma emulsão estável sem qualquer separação de água inicial para os óleos utilizados passou a ser utilizada como condição de agitação padrão para as combinações experimentais de temperatura, salinidade e corte de água definidas pelo LHS. Além disso, foi visado que a combinação propiciasse diâmetro médio de gota em torno de 10 μm, devido a dados de campo conduzirem ao entendimento que uma distribuição centrada nessa media ser representativa (Alves, 2020; Feitosa, 2018).

Definido a melhor combinação de tempo e agitação, para cada uma das combinações experimentais, amostras dos 13 (treze) óleos e da salmoura foram separadas em tubos Falcon de 50 mL e colocadas para aquecimento por 1h em banho termostático transparente ECO ET 15 G Lauda Alpha. Em seguida, cada uma das amostras de óleo e salmoura foram despejadas em beckers de 50mL e misturadas com homogeneizador IKA® T25 digital ULTRA-TURRAX® sob tempo e velocidade de agitação determinados previamente (P1-P9: 3600 rpm / 5 min e P10-P13: 5000 rpm / 5 min). Após o procedimento de agitação para formação das emulsões, as amostras seguiram para o procedimento de caracterização.

3.4 CARACTERIZAÇÃO DE EMULSÕES

Após a preparação, foram realizadas caracterizações das emulsões conforme ilustra o fluxograma da Figura 25.

Figura 25 - Fluxograma da caracterização das emulsões.



Fonte: Próprio autor.

3.4.1 Teste de estabilidade

Seguindo o fluxograma da Figura 25, após o término do processo de mistura da salmoura e do óleo para a formação da emulsão, foi verificado visualmente a ocorrência de separação imediata das fases.

As misturas que apresentaram separação imediata seguiram direto para o acompanhamento de separação em frascos graduados, sendo transferidas dos beckers para frascos cônicos ASTM D91 e colocados em banho termostatizado (Lauda), ajustado para a temperatura desejada conforme as condições da preparação da emulsão, seguindo o planejamento experimental.

A cuba transparente do banho permitiu a observação contínua das emulsões durante todo o experimento, sendo acompanhadas por um período total de 2 horas. Durante este tempo, foi registrada a separação de água na base do frasco a cada intervalo de 5 minutos. O volume

de água separado (quando presente) foi anotado, utilizando a graduação do frasco como referência.

As emulsões que não apresentaram separação imediata, seguiram para a realização de fotomicrografias e teste de viscosidade conforme descrição realizada nas seções 3.4.2 e 3.4.3, respectivamente. Ao fim destas, as amostras foram também transferidas dos beckers para os tubos cônicos ASTM D91 e o comportamento monitorados por 2 horas para verificar sua estabilidade.

As informações foram estruturadas em tabelas, com os volumes de água registrados de acordo com o tempo. Esses registros serviram como base para determinar o percentual de água separada ao fim dos ensaios de estabilidade.

3.4.2 Determinação de distribuição de tamanho e diâmetro médio de gotas

Para determinar a distribuição de tamanho de gota (DTG) de cada uma das emulsões, foi utilizada a técnica de microscopia e análise de imagem. Para isso, das emulsões que apresentaram estabilidade após mistura, alíquotas foram colocadas sobre lâminas (25,4 x 76,2 mm e espessura de 1-1,2 mm) e cobertas com lamínulas de vidro (20 x 20 mm e espessura de 0,13-0,16 mm). Em seguida, as lâminas preparadas foram postas em um suporte com lente objetiva de alta resolução apocromática com aumento de 50 x e abertura de 0,65 com distância focal f de 200 mm, acoplada com câmera digital de 5 MP para que as fotografias fossem realizadas para posterior quantificação das gotas.

A detecção e a quantificação das gotas de água foram realizadas por meio da implementação da transformada de Hough em Python, usando a técnica Hough Circles da biblioteca OpenCV de modo que a quantificação das gotas fornecesse de forma automática a distribuição de tamanho de gota característica da emulsão formulada.

Além da DTG, o algoritmo desenvolvido determinou o diâmetro médio das gotas dispersas no sistema, calculado usando a equação do diâmetro médio de Sauter (28), que relaciona o volume da gota, a área superficial e a contagem de gotas (Jurado *et al.*, 2007).

$$d_{32} = \frac{\sum_{i=1}^{n} d_i^3 \cdot cont_i}{\sum_{i=1}^{n} d_i^2 \cdot cont_i} \tag{28}$$

sendo d_i^2 e d_i^3 os diâmetros de gota de cada intervalo da distribuição elevados ao quadrado e ao cubo respectivamente, e $cont_i$ a contagem de gotas daquele intervalo.

3.4.3 Determinação de viscosidade

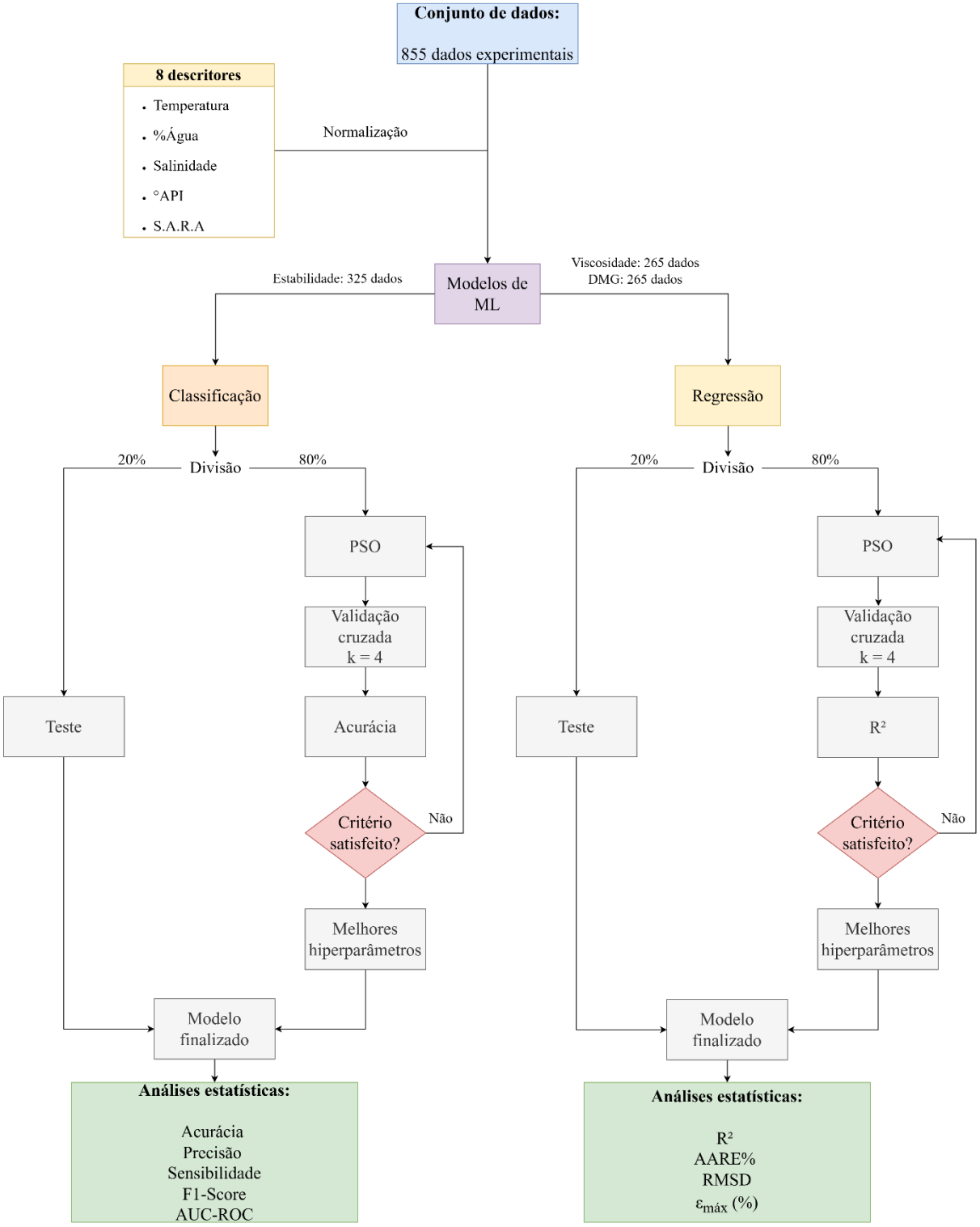
Após a mistura e a retirada de alíquotas para fotomicrografias, as emulsões que não apresentaram separação imediata foram submetidas à análise de viscosidade utilizando um Viscosímetro Rotacional ViscoQC 300 da Anton Paar. Antes de registrar as medições, as amostras foram mantidas no viscosímetro por cerca de 1 minuto, garantindo a estabilização dos valores de viscosidade. Uma vez confirmada a estabilidade das leituras, os valores de viscosidade aparente foram registrados para análises posteriores e o becker com a emulsão seguiram o procedimento descrito na seção 3.4.1.

Para posterior comparação entre a viscosidade de emulsões preditas pelo modelo de aprendizado de máquina e a viscosidade de emulsões calculadas por correlações empíricas clássicas, a viscosidade do óleo foi determinada à pressão atmosférica (0,1 MPa) usando um viscodensímetro Anton Paar SVM 3001 na faixa de temperatura de 313,15 a 353,15 K.

3.5 MODELOS DE APRENDIZADO DE MÁQUINA

Após finalização dos ensaios do planejamento experimental, os dados foram organizados de forma adequada para treinamento dos modelos de aprendizado de máquina. A Figura 26 ilustra a metodologia completa empregada para o desenvolvimento de modelos de aprendizado de máquina a partir do conjunto de dados experimentais.

Figura 26 - Fluxograma de desenvolvimento dos modelos de aprendizado de máquina.

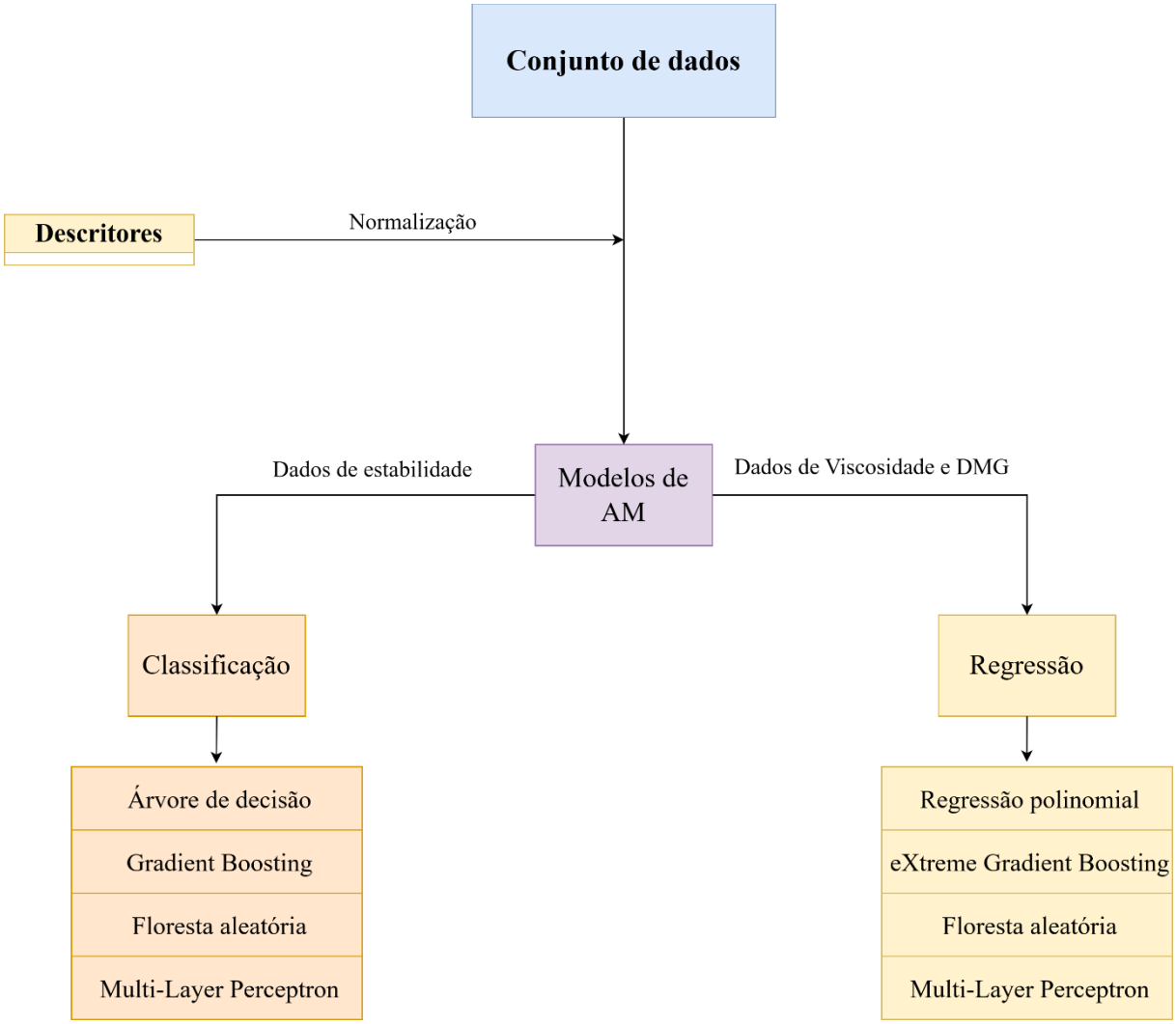


Fonte: Próprio autor.

3.5.1 Modelos de classificação e regressão

Para o desenvolvimento do trabalho, foram utilizados 4 modelos de classificação e 4 modelos de regressão, com mostra a Figura 27.

Figura 27 - Modelos selecionados para classificação e regressão dos dados experimentais.



Fonte: Próprio autor.

A escolha destes algoritmos buscou abranger diferentes abordagens de aprendizado de máquina, incluindo modelos baseados em árvores de decisão, redes neurais e regressão linear estendida. A diversidade de modelos permite comparar o desempenho de algoritmos com diferentes fundamentos teóricos e complexidades, desde modelos mais simples e interpretáveis até *ensembles* e redes neurais.

Para os modelos desenvolvidos, as tabelas do Anexo A2 (Tabelas A.2.1, A.2.2 e A.2.3) reúnem os hiperparâmetros que foram utilizados e otimizados pelo algoritmo PSO para maximizar os resultados de predição.

Os parâmetros utilizados foram selecionados a partir de um teste de sensibilidade, em que foi testado quais parâmetros e a quais limites os modelos apresentavam melhores resultados.

3.5.2 Pré-processamento dos dados

No processo de Aprendizado de Máquina (AM), os parâmetros de entrada frequentemente possuem unidades dimensionais diversas, o que poderia comprometer a precisão e a confiabilidade dos resultados da análise de dados. Uma etapa importante para lidar com esse desafio foi o pré-processamento dos dados. A normalização foi empregada para alinhar esses parâmetros, garantindo uniformidade e reduzindo o potencial viés introduzido pelas discrepâncias dimensionais (Liu *et al.*, 2023). Um intervalo entre 0,1 e 0,9 foi utilizado para a normalização das variáveis de entrada (Khataee; Kasiri, 2010; Sousa *et al.*, 2014), conforme Equação (29):

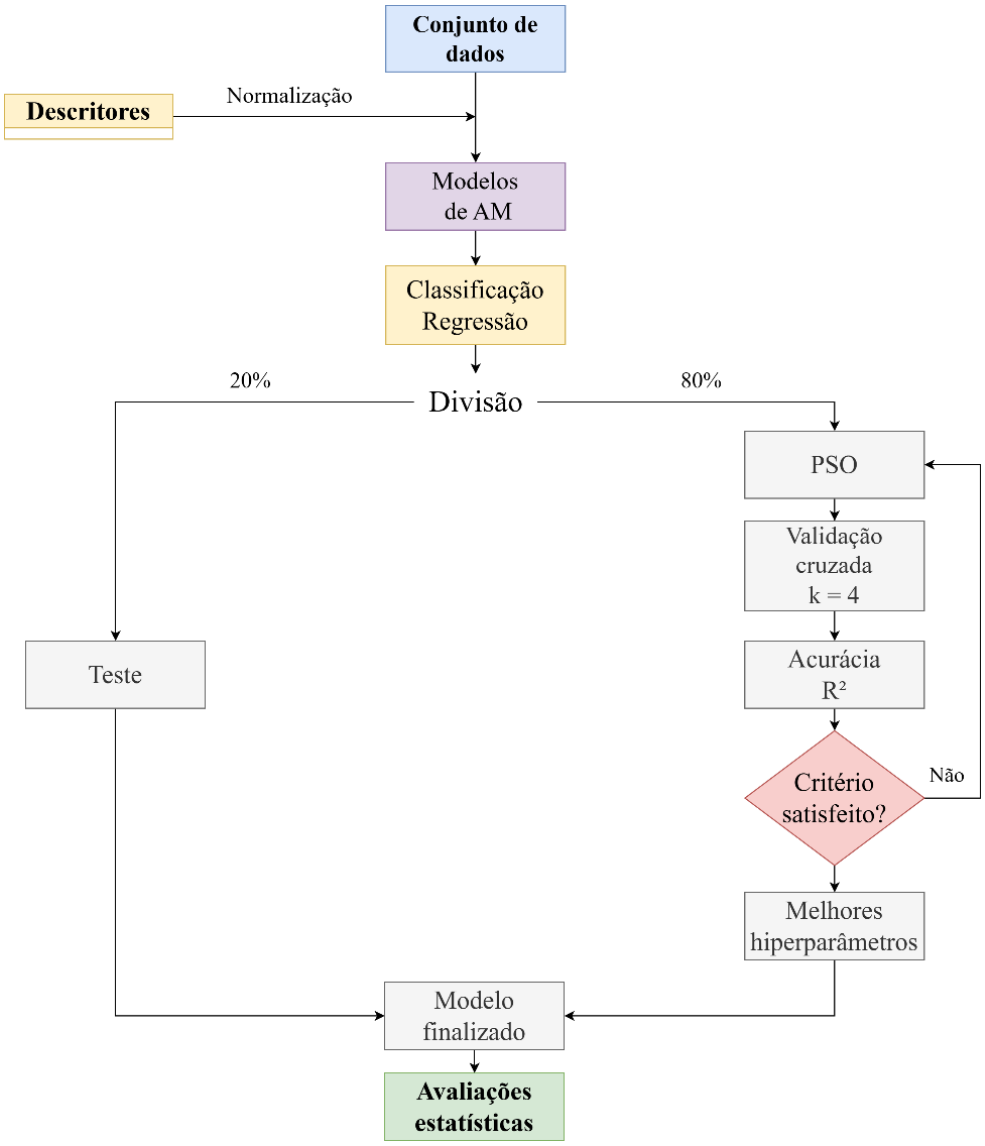
$$\theta' = 0{,}8\left(\frac{\theta - \theta_{min}}{\theta_{max} - \theta_{min}}\right) + 0{,}1 \tag{29}$$

onde θ' representa a variável normalizada, e os subscritos min e max são os valores mínimo e máximo da amostra, respectivamente.

3.5.3 Treinamento e otimização de hiperparâmetros

A partir do fluxograma geral de desenvolvimento dos modelos da Figura 26, a etapa de treinamento e otimização dos hiperparâmetros seguiu o fluxograma da Figura 28.

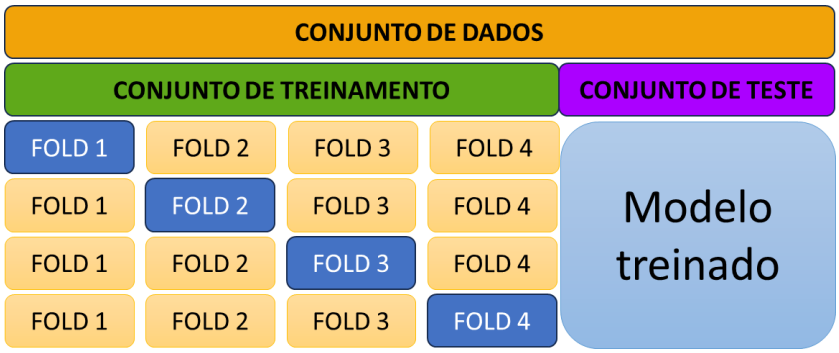
Figura 28 - Fluxograma de desenvolvimento e otimização dos modelos.



Fonte: Próprio autor.

Para otimizar os modelos de aprendizado de máquina, foi implementada a Otimização por Enxame de Partículas (PSO) com o objetivo de encontrar os hiperparâmetros ideais. A avaliação do desempenho de cada conjunto de hiperparâmetros foi guiada por funções objetivo específicas ao tipo de problema: Acurácia para modelos de classificação e o coeficiente de determinação (R^2) para modelos de regressão. A robustez dessa avaliação e a prevenção do sobreajuste (*overfitting*) foram garantidas pelo uso da validação cruzada k-*fold* (k = 4), conforme ilustrado na Figura 29, aplicada ao conjunto de treinamento.

Figura 29 - Diagrama do processo de validação cruzada k-fold.



Fonte: Próprio autor.

O algoritmo PSO foi configurado para operar com uma população de 100 partículas e um máximo de 100 iterações. Os critérios de parada para o processo de otimização incluíram tanto o atingimento desse número máximo de iterações quanto a convergência da solução – esta última definida como uma melhoria na melhor combinação de hiperparâmetros inferior a 1×10^{-8} em relação às iterações anteriores. Ao final do processo, o PSO identificava os hiperparâmetros considerados ótimos, os quais foram documentados em anexo (Tabelas A.2.1, A.2.2 e A.2.3).

Posteriormente, utilizando esses hiperparâmetros otimizados, um modelo final foi treinado. Por último, para assegurar a validade e a capacidade de aplicação do modelo em cenários gerais, seu desempenho e capacidade de generalização foram avaliados em um conjunto de teste distinto (não utilizado em nenhuma etapa anterior de treinamento ou seleção de hiperparâmetros) e as predições analisadas estatisticamente.

3.6 AVALIAÇÃO DAS PREDIÇÔES DOS MODELOS

A avaliação do desempenho de modelos é uma etapa crucial no desenvolvimento de soluções analíticas, pois garante que os resultados sejam confiáveis, precisos e aplicáveis a situações reais. Sem uma análise rigorosa, um modelo pode apresentar falhas como *overfitting* (quando se ajusta excessivamente aos dados de treinamento) ou *underfitting* (quando não captura padrões relevantes), comprometendo sua capacidade de generalização para novos dados (Géron, 2022).

A escolha de métricas adequadas permite identificar pontos fortes e fracos do modelo, orientando ajustes que alinhem suas previsões aos objetivos do problema em questão, seja

maximizar a acurácia, minimizar erros críticos ou atender a demandas específicas de um domínio, como saúde ou finanças (Alpaydin, 2020).

3.6.1 Modelos de classificação

Para avaliar o desempenho dos modelos de classificação desenvolvidos neste trabalho, foram selecionadas métricas amplamente utilizadas na literatura, que permitem analisar diferentes aspectos da performance preditiva. Essas métricas foram escolhidas com base em sua capacidade de fornecer uma visão abrangente sobre a qualidade das predições, considerando tanto os acertos quanto os erros do modelo.

3.6.1.1 Avaliação estatística

Em um cenário de classificação binária, 4 conceitos são fundamentais para realização dos cálculos, são eles: Verdadeiros Positivos (VP), Verdadeiros Negativos (VN), Falsos Positivos (FP) e Falsos Negativos FN). VP e VN são as predições corretas, enquanto FP e FN representam os erros. Esses conceitos foram utilizados para estimar a acurácia, precisão, recall/sensibilidade e F1-score dos modelos treinados.

- **Acurácia:** representa a proporção de predições corretas realizadas pelo modelo em relação ao total de amostras avaliadas

$$Acurácia = \frac{VP + VN}{VP + VN + FP + FN} \tag{30}$$

Apesar de ser uma métrica intuitiva, seu uso foi complementado por outras métricas para lidar com possíveis desbalanceamentos no conjunto de dados.

- **Precisão:** mede a proporção de predições positivas corretas em relação ao total de predições positivas feitas pelo modelo. É especialmente útil em cenários onde o custo de falsos positivos é elevado, pois avalia a confiabilidade das predições positivas sendo. definida como:

$$Precisão = \frac{VP}{VP + FP} \tag{31}$$

- **Recall ou Sensibilidade:** calcula a proporção de casos positivos corretamente identificados pelo modelo em relação ao total de casos positivos reais, expressa por:

$$Recall/Sensibilidade = \frac{VP}{VP + FN} \tag{32}$$

- **F1-Score:** É a média harmônica entre *Precisão* e S*ensibilidade*, fornecendo um equilíbrio entre essas duas métricas. Essa métrica foi utilizada para avaliar o desempenho geral do modelo, especialmente em situações de classes desbalanceadas, onde a acurácia isolada poderia ser enganosa. É calculado como:

$$F1 - Score = 2 \cdot \frac{Precisão \cdot Sensibilidade}{Precisão + Sensibilidade} \tag{33}$$

- AUC-ROC: É a área abaixo das curvas (*Area Under Curve* - AUC) características de operação do receptor (*Receiver Operating Characteristic* - ROC). Essa métrica é utilizada como avaliação do desempenho de modelões de classificação binária.

3.6.1.2 Avaliação de desempenho

A avaliação de desempenho foi realizada determinando-se a área abaixo das curvas características de operação do receptor (AUC-ROC). A curva ROC é uma representação gráfica que demonstra o desempenho de um sistema de classificação binária em vários limiares. Ela é construída com base na taxa de verdadeiros positivos (TVP) e na taxa de falsos positivos (TFP) dos classificadores analisados (Arjaria; Rathore; Cherian, 2021).

Para calcular as curvas ROC de cada um dos modelos, os limiares (ou *thresholds*) foram ajustados gradativamente, começando de valores altos (próximos de 1) até valores baixos (próximos de 0). A cada ajuste, o modelo classificou as emulsões como estáveis ou instáveis, e os rótulos preditos foram comparados com os rótulos reais. A partir dessas comparações, foram calculados valores de VP, FP, VN e FN. Com esses valores, a TVP e a TFP foram determinadas para cada *threshold*, calculadas pelas equações (21) e (23) respectivamente.

Os valores de TVP e TFP calculados para cada *threshold* formaram os pontos da curva ROC. A área sob a curva (AUC) sintetiza essa análise em uma única métrica, onde valores mais

altos de AUC indicam melhor capacidade do modelo para discriminar entre as classes positiva e negativa em todos os *thresholds* avaliados.

3.6.2 Modelos de regressão

Para avaliar a eficácia dos modelos de aprendizado de máquina, foram utilizadas as seguintes métricas.

- Coeficiente de determinação (R^2): O coeficiente de determinação é uma métrica estatística amplamente utilizada para avaliar quão bem os valores previstos pelo modelo correspondem aos valores experimentais ou observados.

$$R^2 = 1 - \left(\frac{\sum_{i=1}^{N} \left(y_i^{exp} - y_i^{pred} \right)^2}{\sum_{i=1}^{N} \left(y_i^{exp} - y_i^{\overline{exp}} \right)^2} \right) \tag{34}$$

O valor obtido a partir desse cálculo indica o grau de variabilidade dos dados experimentais que é explicado pelos valores previstos pelo modelo. Quanto mais próximo de 1 for o resultado, melhor será o desempenho do modelo na explicação da variação observada (Belyadi; Haghighat, 2021).

- AARE%: é uma métrica usada para medir quão preciso é um modelo ou medição. Ele calcula a média dos erros relativos absolutos entre os valores previstos e os reais, geralmente expressa como uma porcentagem, ajudando a entender a magnitude média dos erros em relação aos valores reais.

$$AARE\% = \frac{100}{N} \sum_{i=1}^{N} \left| \frac{y_i^{exp} - y_i^{pred}}{y_i^{exp}} \right| \tag{35}$$

- Raiz do desvio quadrático médio (RMSD – *Root Mean Square Deviation*): Uma métrica de desempenho típica para problemas de regressão é o RMSD. Ele dá uma ideia de quanto erro o sistema geralmente comete em suas previsões, atribuindo um peso maior para erros grandes (Géron, 2022). A Equação (36) apresenta a fórmula matemática para calcular o RMSD.

$$RMSD = \sqrt{\frac{100}{N}\sum_{i=1}^{N}\left(y_i^{exp} - y_i^{pred}\right)^2} \quad (36)$$

O RMSD começa com o cálculo do desvio $e_i = y_i^{exp} - y_i^{pred}$, que representa a diferença entre o valor real e a previsão do modelo.

- $\varepsilon_{máx}$ (%): O erro absoluto máximo relativo é a maior diferença absoluta entre os valores previstos por um modelo e os valores observados. Ele indica o pior caso de desvio, sendo útil para identificar onde o modelo ou medição falha mais significativamente.

3.6.3 Discriminação entre modelos

Para discriminar e ordenar os modelos por eficácia, foi aplicado o teste de Friedman. Este é um método não paramétrico usado como alternativa à ANOVA de medidas repetidas, que avalia o desempenho dos algoritmos, classificando-os para cada conjunto de dados. Ao algoritmo com melhor desempenho atribui-se o posto 1, ao segundo melhor, o posto 2, e assim por diante. Em caso de empates, são atribuídas classificações médias aos algoritmos correspondentes (Ma *et al.*, 2022). O teste de Friedman avalia as seguintes hipóteses:

1. Hipótese nula (H_0): todos os métodos comparados apresentam desempenho equivalente, ou seja, não há diferença significativa entre suas medianas.
2. Hipótese alternativa (H_1): indica que pelo menos um dos métodos difere significativamente dos demais.

O valor da estatística de teste é calculado conforme a Equação (37):

$$Q = \frac{12}{nk(k+1)}\sum_{j=1}^{k} R_j^2 - 3n(k+1) \quad (37)$$

em que n é o número de casos avaliados, k é o número de métodos comparados, e R_j é a soma das classificações (rankings) do j-ésimo método. A hipótese nula é rejeitada quando o valor de

Q calculado excede o valor crítico da distribuição qui-quadrado com *k – 1* graus de liberdade, ao nível de significância escolhido (α = 0,05).

O teste pós-hoc de Nemenyi foi aplicado para identificar quais pares de modelos apresentaram diferenças estatisticamente significativas. Semelhante ao teste de Tukey para ANOVA, o teste de Nemenyi é utilizado quando todos os modelos são comparados entre si. Uma diferença de desempenho considerável entre dois modelos é observada quando suas classificações médias diferem, no mínimo, pela diferença crítica, conforme mostrado na Equação (38).

$$DC = q_{\alpha}\sqrt{\frac{k(k+1)}{6n}} \tag{38}$$

em que q_{α} são baseados na estatística de amplitude studentizada dividida por $\sqrt{2}$ (Demšar, 2006).

Duas classificações médias são consideradas significativamente diferentes quando sua diferença excede a diferença crítica (DC); caso contrário, os modelos são considerados estatisticamente semelhantes. Os resultados do teste de Nemenyi podem ser apresentados em uma matriz de p-valores ou em diagramas de diferença crítica, que indicam visualmente os grupos de modelos com desempenho equivalente. Os testes de Friedman e Nemenyi foram realizados utilizando funções implementadas das bibliotecas SciPy e scikit_posthocs, em Python (Santos *et al.*, 2024; Terpilowski, 2019; Virtanen *et al.*, 2020).

3.7 ANÁLISE SHAP DOS PARÂMETROS DE ENTRADA DOS MODELOS

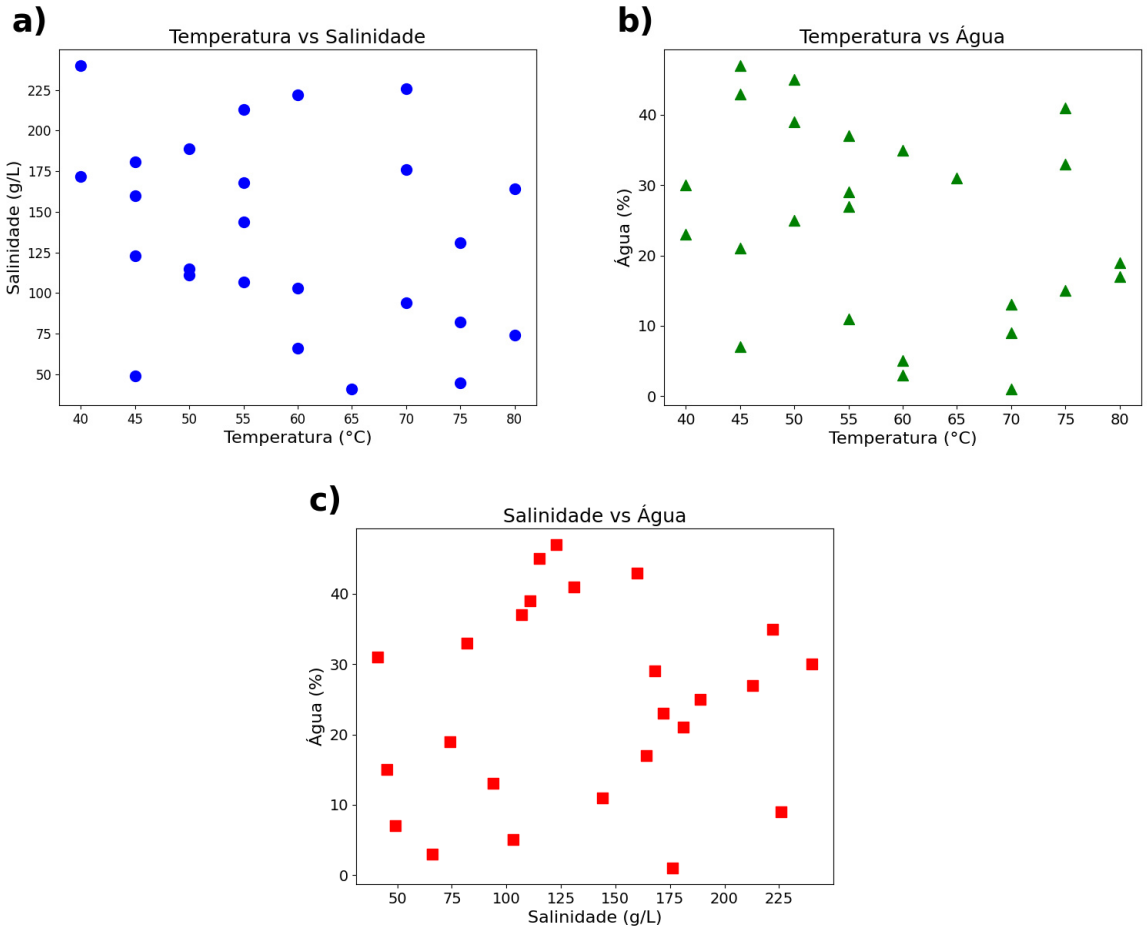
A metodologia SHAP (*SHapley Additive exPlanations*) foi aplicada para avaliar o impacto dos parâmetros de entrada na previsão de propriedades. Essa análise permitiu identificar a contribuição dos parâmetros de entrada, auxiliando na compreensão de como elas influenciam os resultados previstos pelo modelo. O método SHAP atribui uma pontuação de importância a cada característica, quantificando seu impacto nas saídas e representando seu papel no processo preditivo (Meng *et al.*, 2023).

4 RESULTADOS

4.1 PLANEJAMENTO EXPERIMENTAL COM LHS

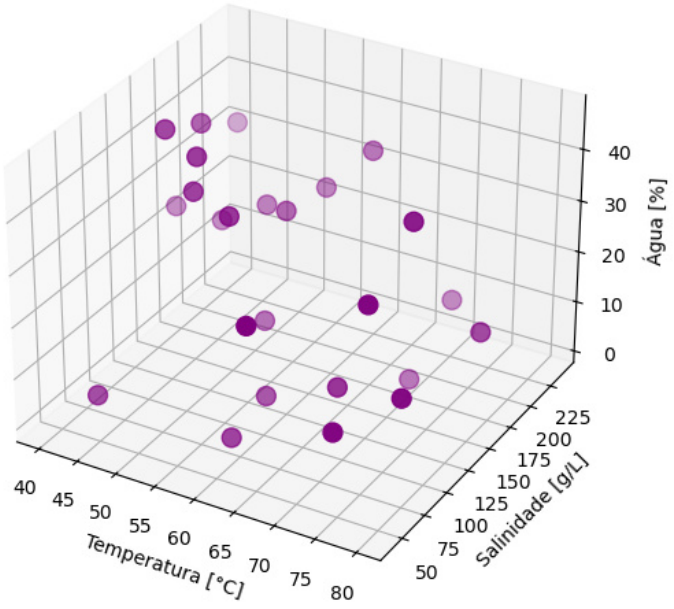
O método LHS foi empregado com êxito na otimização do planejamento experimental considerando como parâmetros a temperatura, a salinidade e o corte de água para a formulação de emulsões de petróleo, permitindo a geração de 25 combinações experimentais que exploram de maneira eficiente o espaço paramétrico definido. A Figura 30 e a Figura 31 mostram a visualização da distribuição dos pontos experimentais nos espaços 2D e 3D, respectivamente.

Figura 30 - Distribuição dos pontos em uma perspectiva bidimensional (2D) para cada variável de entrada.



Fonte: Próprio autor.

Figura 31 - Distribuição dos pontos experimentais em uma perspectiva tridimensional (3D).



Fonte: Próprio autor.

O uso do LHS garantiu uma distribuição representativa dos pontos experimentais ao longo das faixas estabelecidas, minimizando o número de experimentos necessários enquanto maximiza a cobertura do espaço de projeto. Esse método se mostrou particularmente vantajoso em um contexto de formulação de emulsões, onde interações complexas entre os parâmetros podem influenciar significativamente as propriedades finais do sistema.

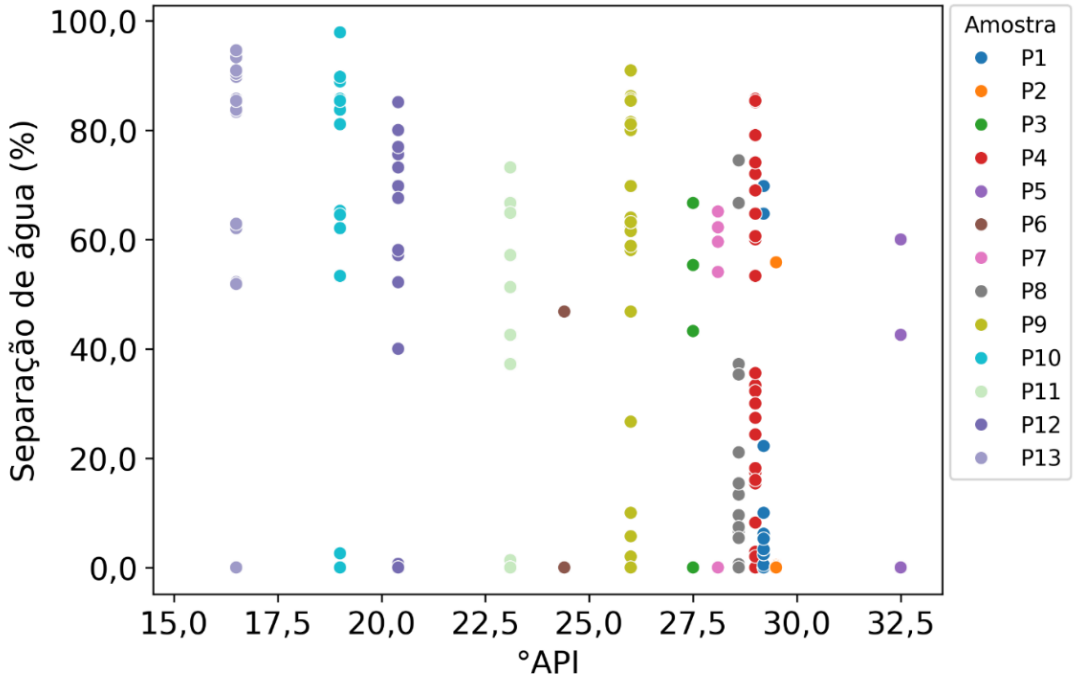
Os espaços vazios visíveis nas plotagens de pares de variáveis são uma consequência do LHS. Em vez de testar todas as combinações possíveis, o método foca em garantir que a projeção dos pontos em cada eixo cubra toda a faixa daquela variável. Na prática, isso significa que cada parâmetro foi testado em seus diversos níveis (baixo, médio, alto), obtendo a máxima informação sobre o efeito de cada fator com um número reduzido de experimentos.

4.2 CARACTERIZAÇÃO DOS ENSAIOS

4.2.1 Separação de água e viscosidade das emulsões

Foram realizados um total de 325 (trezentos e vinte e cinco) testes de formação de emulsões. A Figura 32 mostra os resultados de separação de água ao final do processo de monitoramento da estabilidade para cada um dos óleos utilizados.

Figura 32 - Separação de água em função do °API dos óleos.

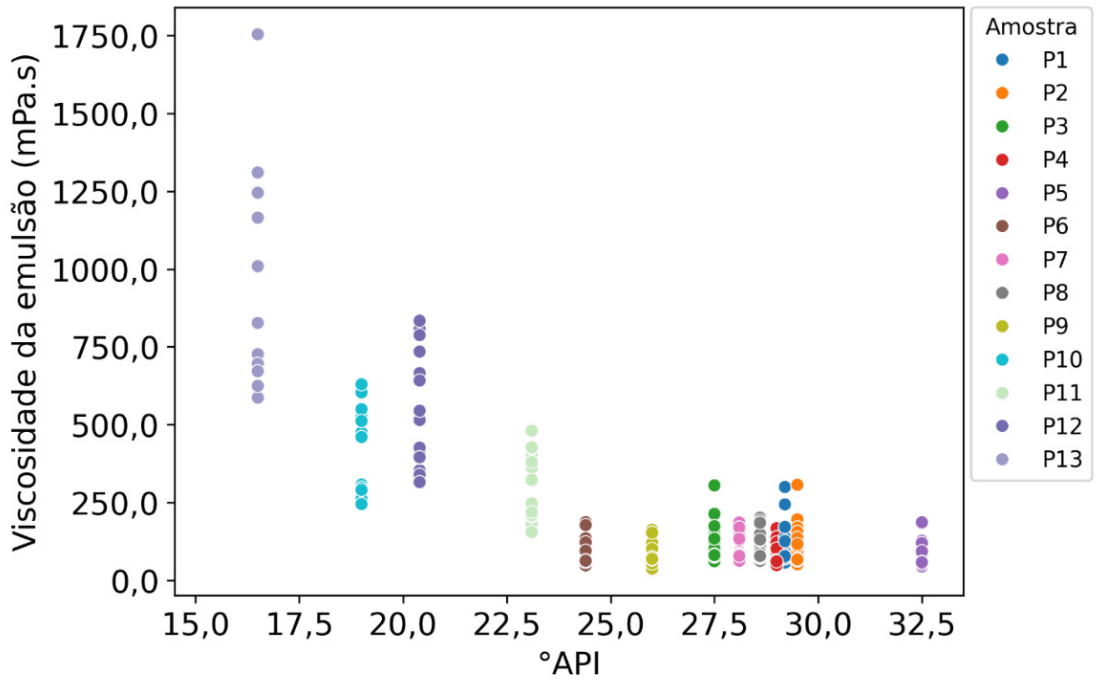


Fonte: Próprio autor.

Os resultados mostram variabilidade na separação de água para cada um dos óleos utilizados, com amostras de menor °API apresentando maiores graus de separação de água. No entanto, essa tendência não é uniforme, pois óleos com °API mais alto, como P4 e P9, também exibiram separações significativas, alcançando valores acima de 80%. Isso indica que, além do °API, outros fatores como a composição química do óleo e as diferentes condições experimentais proporcionada pelo método LHS desempenham influência na estabilidade das emulsões, acarretando em diferentes graus de separação de água do óleo.

Além dos resultados de separação de água, a Figura 33 mostra os resultados experimentais para viscosidade de emulsões para cada um dos óleos utilizados nas diferentes condições do método LHS.

Figura 33 - Viscosidade da em função do °API dos óleos.



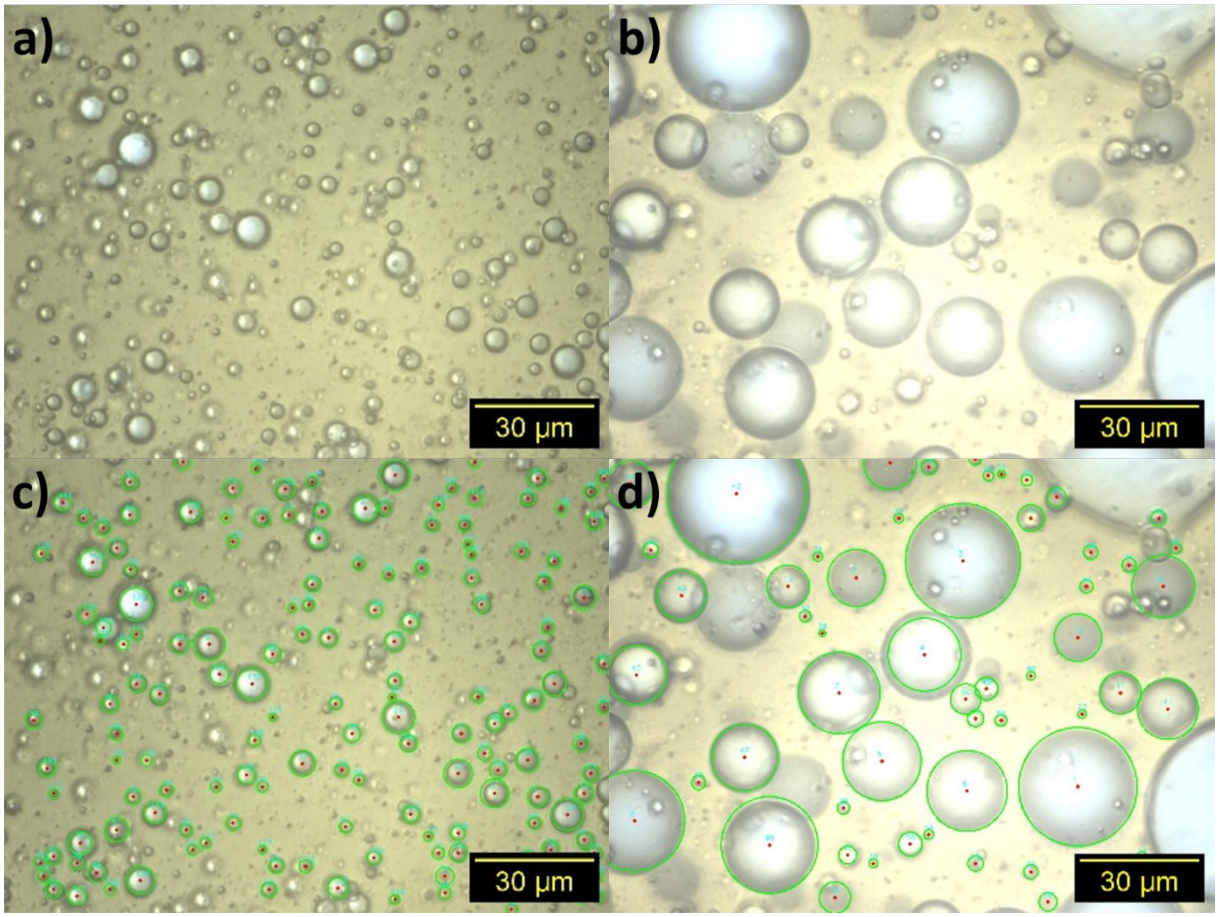
Fonte: Próprio autor.

O gráfico mostra a relação entre a viscosidade da emulsão e a gravidade API (°API) para as diferentes amostras de óleo. Nota-se uma tendência geral, em que à medida que a gravidade API aumenta, a viscosidade da emulsão tende a diminuir. Para viscosidades acima de 250 mPa.s há menos densidade de pontos experimentais quando comparado com a densidade de pontos abaixo dessa faixa, podendo ser um desafio no desenvolvimento dos modelos de aprendizado de máquina.

4.2.2 Determinação do diâmetro médio de gota a partir das fotomicrografias

Após finalização do monitoramento para verificação de estabilidade das misturas, as fotomicrografias realizadas das amostras que apresentaram estabilidade inicial foram tratadas para determinação das distribuições de tamanho de gota e o tamanho médio das gotas. A Figura 34 mostra fotomicrografias de duas amostras em diferentes condições experimentais (ensaios 4 e 22 do anexo A.1 para as amostras P10 e P2, respectivamente) antes e após a aplicação do algoritmo com a técnica *HoughCircles* para detecção e quantificação das gotas, ferramenta desenvolvida no trabalho.

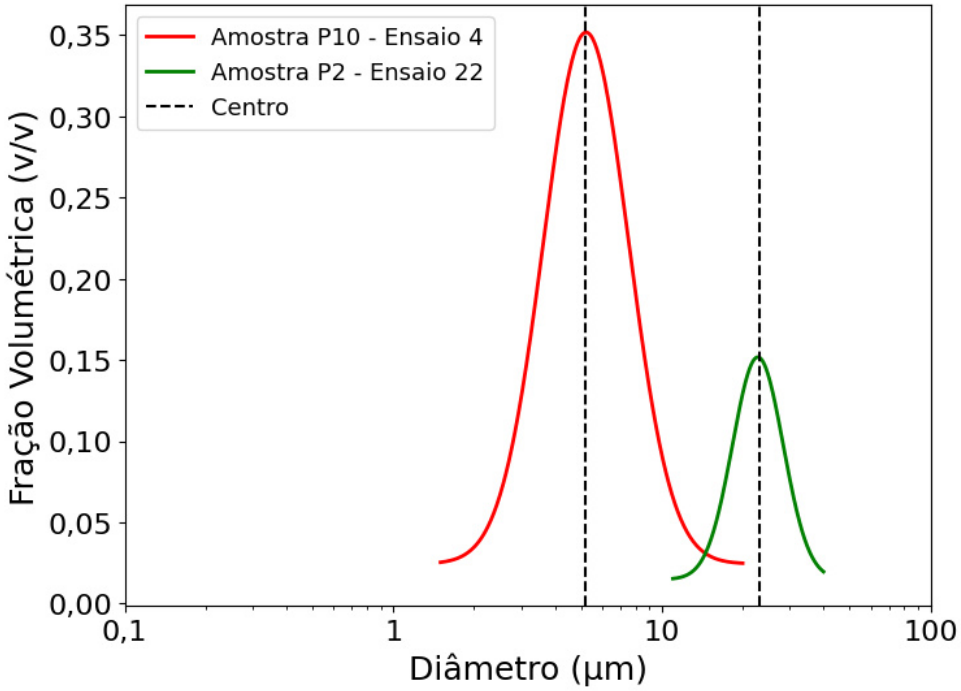
Figura 34 - Fotomicrografias de 2 amostras antes da aplicação da técnica HC em a) e b) e as mesmas após aplicação da técnica em c) e d).



Fonte: Próprio autor.

Os resultados da aplicação da técnica *HoughCircles* mostraram-se satisfatórias, detectando com precisão os contornos das gotas de água dispersas no meio contínuo. Mesmo em condições variadas, como %água (Ensaio 4 - 7% e Ensaio 22 - 33%) que implicam em alterações no tamanho das gotas e variações na luminosidade que podem dificultar a detecção. A partir da quantificação das gotas, foi calculado o diâmetro médio de gota usando a Equação (28). Curvas de fração volumétrica em função do diâmetro para amostras apresentadas na Figura 34 são mostrados na Figura 35.

Figura 35 - Distribuição do tamanho de gotas para 2 amostras: amostra P10 do ensaio 4 e amostra P2 do ensaio 22.

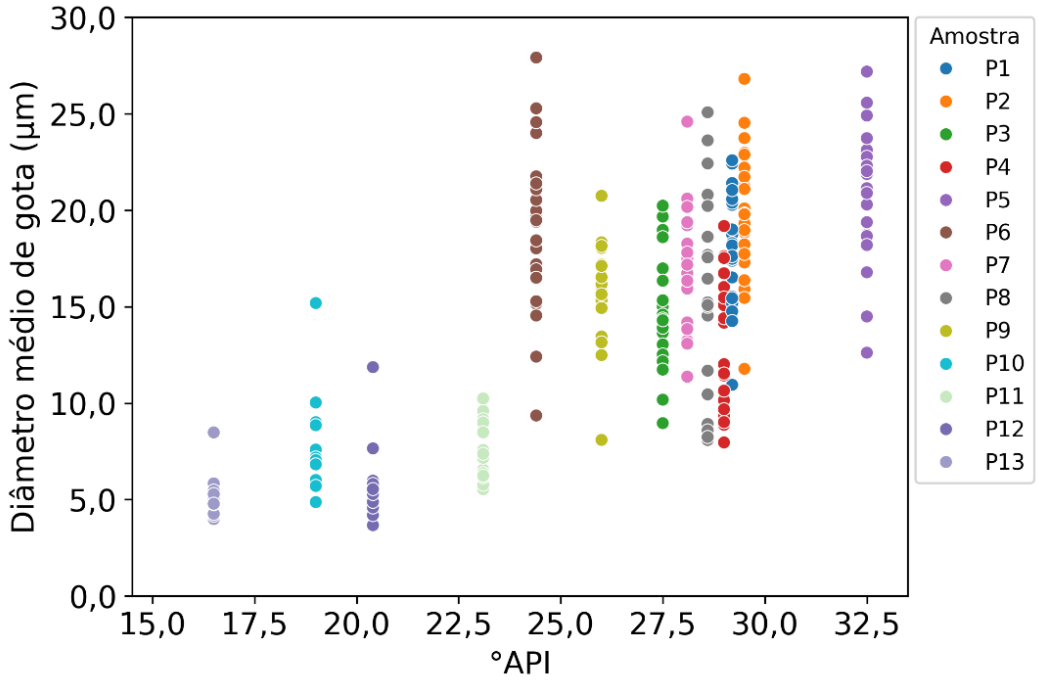


Fonte: Próprio autor.

A análise comparativa das distribuições de tamanho de gotas revela uma diferença acentuada entre as amostras, com a Amostra P10 (Ensaio 4 - 7% de água) exibindo um diâmetro significativamente menor em contraste com a Amostra P2 (Ensaio 22 - 33% de água), que apresenta gotas consideravelmente maiores. Esta variação sugere que as condições experimentais de cada ensaio impactaram diretamente a morfologia final da dispersão. A hipótese central para explicar tal disparidade é a diferença no corte de água entre os ensaios, que teria influenciado o nível de interação e a probabilidade de coalescência entre as gotas.

A partir da quantificação das gotas, foi calculado o diâmetro médio de gota usando a Equação (28), implementada no algoritomo da ferramenta de detecção de gotas a partir das micrografias. Os resultados de DMG para as amostras da são mostrados na Figura 36.

Figura 36 - Resultados de diâmetro médio de gota em função do °API.

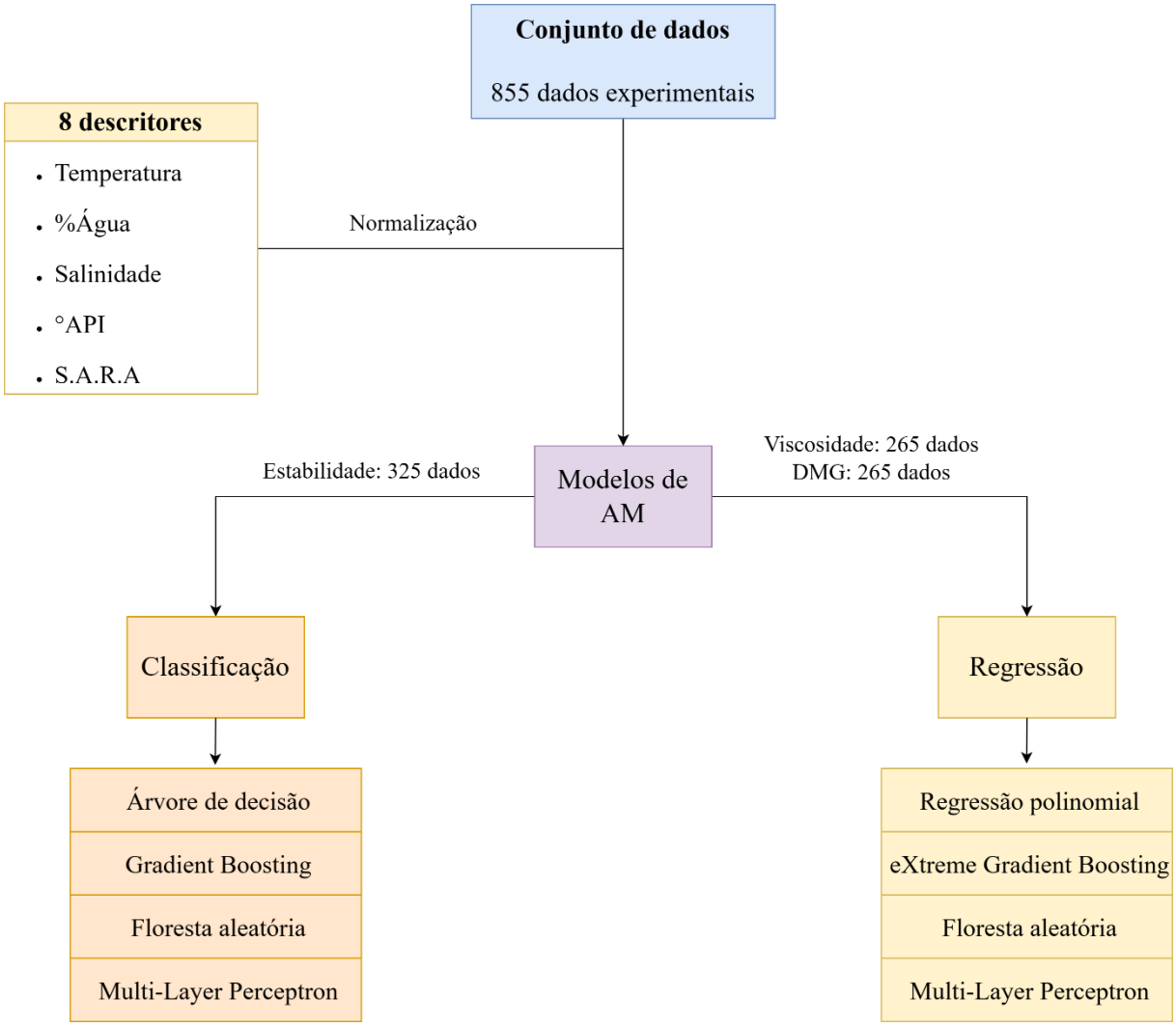


Fonte: Próprio autor.

Os resultados para DMG em função da viscosidade estão condizentes com literatura, em que quanto maior a viscosidade da fase contínua (menor valor de °API), maior a resistência das gotas de água à colisão e à coalescência, dificultando a formação de gotas maiores, enquanto que menores viscosidades facilitam a interação e a colisão das gotas de água (Lv *et al.*, 2024). Finalizado a caracterização das amostras e organização dos dados, foi iniciado o processo de desenvolvimento dos modelos de aprendizado de máquina para realizar as predições.

4.3 MODELOS DE APRENDIZADO DE MÁQUINA

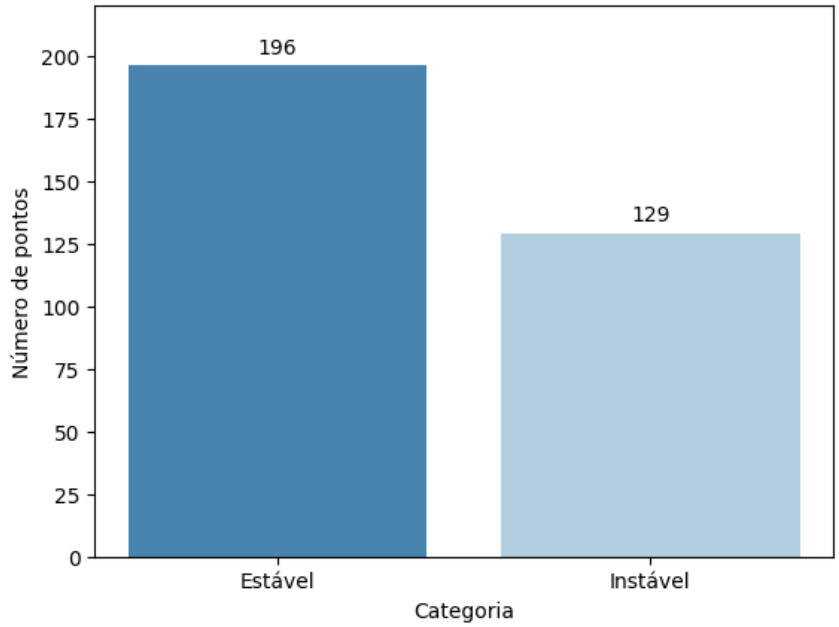
Tomando por base informações da literatura dos fatores que afetam a estabilidade das emulsões, a viscosidade e o diâmetro médio das gotas, além dos 3 descritores variados pelo método LHS, propriedades do petróleo foram utilizadas como variáveis de entrada para os modelos. Assim, as entradas para os modelos de aprendizado de máquina foram: a temperatura, a salinidade, o corte de água, o °API e a composição SARA de cada um dos óleos.



4.3.1 Estabilidade de emulsões

Para realizar predições da estabilidade de emulsões, os modelos *Árvore de Decisão* (AD), *Gradient Boosting* (GB), *Floresta Aleatória* (FA) e *Multi-Layer Perceptron* (MLP) foram utilizados. A classificação de estabilidade teve como critério o percentual de separação de água no teste de acompanhamento de 2 horas. Além das amostras que já apresentaram separação visual de água no tempo 0 após mistura, as que apresentaram separação ao longo do acompanhamento até o fim das 2 horas foram categorizadas como instáveis. A Figura 37 mostra a contagem de emulsões, categorizadas de acordo com sua estabilidade.

Figura 37 - Contagem de emulsões estáveis e instáveis.

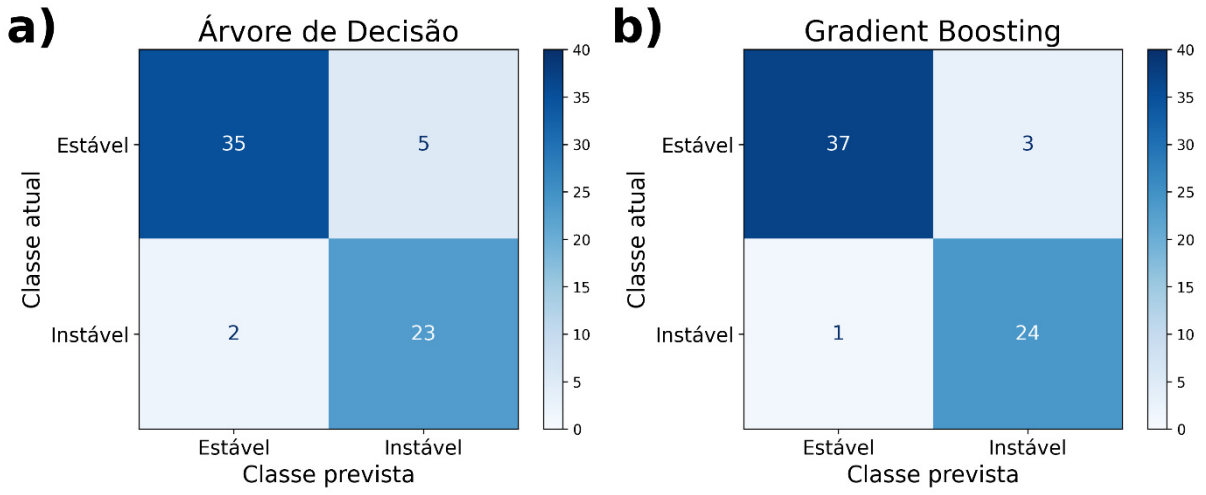


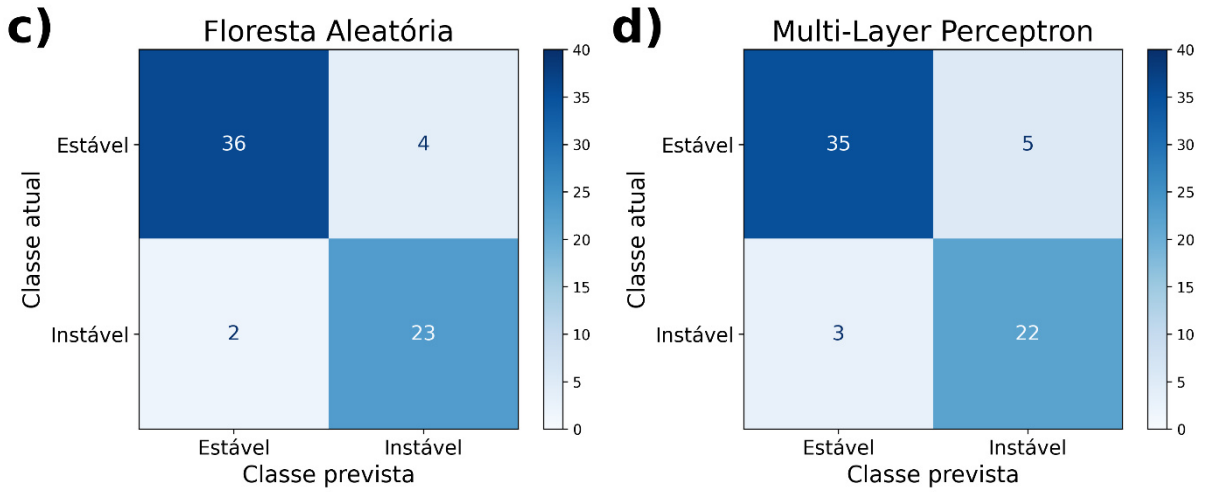
Fonte: Próprio autor.

As métricas calculadas no conjunto de treino e de teste fornecem uma base quantitativa para avaliar os modelos. A Figura 38 apresenta a matriz de confusão para o conjunto de teste dos 4 modelos de classificação testados.

Na matriz de confusão, os valores da diagonal principal representam o número de amostras de cada classe que foram corretamente classificadas. Em contrapartida, os valores da diagonal secundária indicam as classificações incorretas, destacando os erros cometidos pelo modelo (Shakouri; Mohammadzadeh-Shirazi, 2025). As métricas estatísticas para o conjunto de teste e treinamento foram dispostas na Tabela 3.

Figura 38 - Matrizes de confusão para os modelos de classificação a) AD, b) GB, c) FA e d) MLP.





Fonte: Próprio autor.

Tabela 3 - Métricas estatísticas para os modelos de classificação.

Modelos	Métricas									
	Acurácia		Precisão		Sensibilidade		F1-Score		AUC-ROC	
	Treino	Teste	Treino	Teste	Treino	Teste	Treino	Teste	Treino	Teste
AD	0,896	0,892	0,933	0,946	0,891	0,875	0,911	0,909	0,973	0,906
GB	**0,985**	**0,938**	**0,981**	**0,974**	**0,994**	**0,925**	**0,987**	**0,949**	**0,986**	**0,945**
FA	0,973	0,908	0,975	0,947	0,981	0,900	0,978	0,923	0,997	0,925
MLP	0,927	0,877	0,936	0,921	0,942	0,875	0,939	0,897	0,979	0,918

Fonte: Próprio autor.

A partir da Figura 38 e Tabela 3, conclui-se que o modelo MLP apresentou desempenho inferior em comparação aos demais. Sua acurácia no conjunto de teste (0,877) indica uma eficácia aceitável na classificação, embora seja inferior aos demais modelos. Além disso, a precisão (0,921) e sensibilidade (0,875) refletem um equilíbrio razoável, especialmente na identificação de emulsões estáveis e instáveis, mas seu F1-Score de 0,897 revela uma performance geral que ainda requer ajustes. Seu desempenho menos expressivo no conjunto de teste sugere que o modelo pode estar enfrentando desafios em generalizar predições para novos dados, possivelmente necessitando de otimização de mais hiperparâmetros para alcançar melhor desempenho, ou ampliar o espaço de busca.

Por outro lado, os modelos baseados em árvores, incluindo AD, FA e GB mostraram resultados significativamente melhores. O modelo AD teve um desempenho consistente entre os conjuntos de treino e teste, com *Acurácia* de 0,892 no teste. Seus resultados para *Precisão* (0,946) e *Sensibilidade* (0,875) indicam sua confiabilidade. O *F1-Score* de 0,909 confirma sua capacidade equilibrada de lidar com as diferentes classes. Apesar de não ser o modelo com os

melhores resultados, sua simplicidade e interpretabilidade tornam a Árvore de Decisão uma escolha viável. Contudo, o modelo FA superou a AD em todas as métricas, com acurácia de 0,908 e um F1-Score elevado de 0,923 no conjunto de teste, indicando maior eficiência nas classificações.

O modelo GB destacou-se por atingir os melhores resultados entre todas as métricas avaliadas, com *Acurácia* de 0,938, *Precisão* de 0,974, *Sensibilidade* de 0,925, *F1-Score* de 0,949 e AUC de 0,945. Sua superioridade pode ser atribuída à estrutura de *gradient boosting*, que corrige erros sequencialmente, e à utilização de parâmetros de regularização (como a taxa de aprendizado e a profundidade máxima das árvores) que auxiliam no controle do sobreajuste.

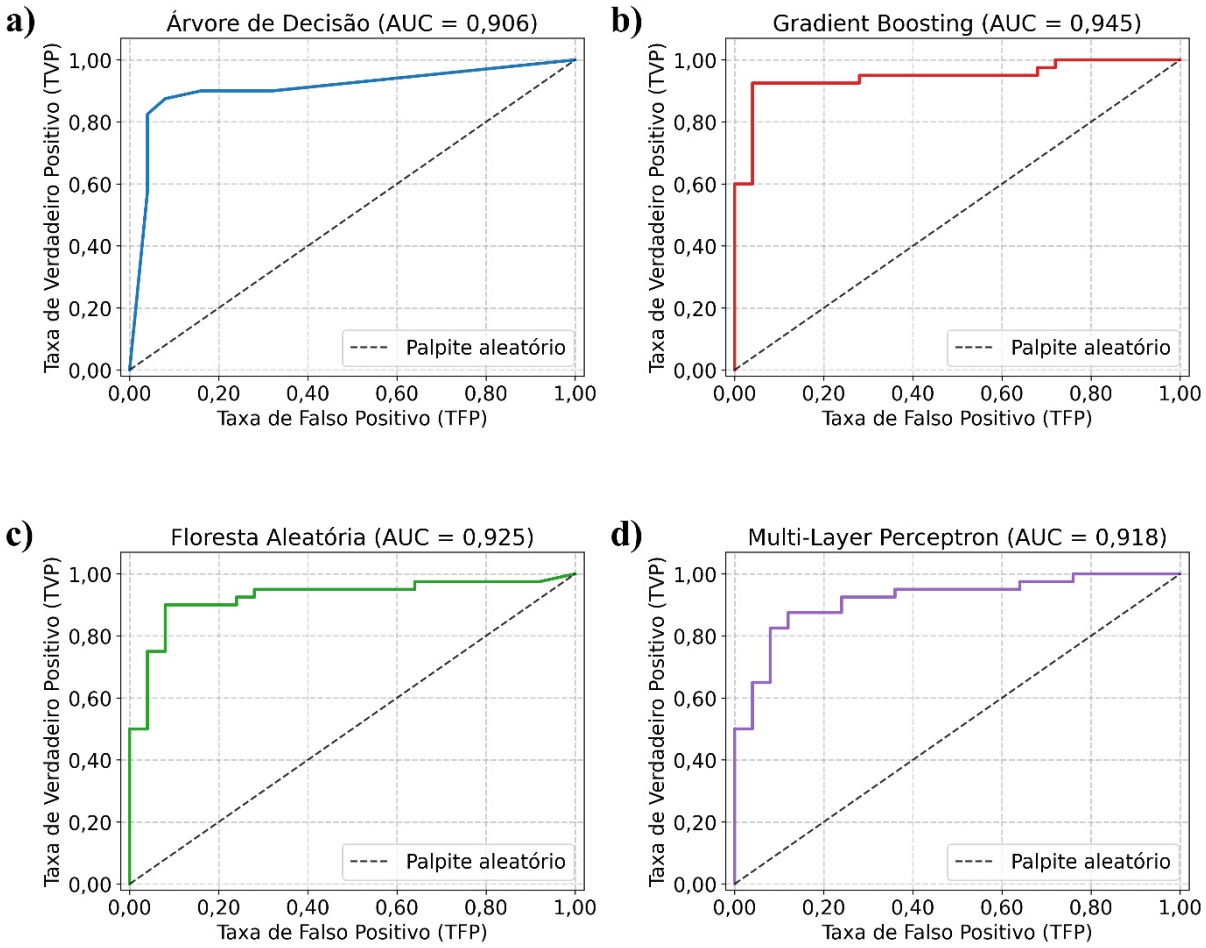
No contexto de classificação de emulsões de petróleo, é essencial priorizar o recall/sensibilidade para minimizar erros críticos. Classificar uma emulsão instável como estável pode causar separação indesejada e custos operacionais elevados, afetando processos como transporte e refino, enquanto classificá-las incorretamente como estáveis pode levar a custos operacionais elevados. Modelos com alta sensibilidade, como GB e FA (0,925 e 0,900, respectivamente), ajudam a evitar falsos negativos. Além disso, o F1-score mais alto do GB (0,949 contra 0,923 da FA) sugere um melhor equilíbrio entre precisão e sensibilidade, tornando-o potencialmente mais adequado.

4.3.1.1 Avaliação de desempenho

Para avaliação de desempenho dos modelos (AD, GB, FA e MLP), foi utilizado curvas ROC para comparar seus respectivos desempenhos em diferentes limiares de corte. O limiar, ou *threshold*, é o ponto de corte usado para decidir se uma amostra pertence à classe positiva ou negativa (emulsão estável ou instável), baseado nas probabilidades preditas pelo modelo.

Um *threshold* padrão geralmente é 0,5, mas é possível ajustá-lo de acordo com as necessidades específicas do problema, alterando o comportamento do modelo. A Figura 39 apresenta as curvas ROC dos modelos de classificação feito varredura em diferentes *thresholds* no intervalo de 0 até 1 e suas respectivas AUC.

Figura 39 - Curvas ROC e AUC dos modelos de classificação.



Fonte: Próprio autor.

O valor AUC expressa o quão bem o modelo separa entre as duas classes, valores mais altos apontam para uma menor taxa de erros na classificação. Para o conjunto de teste, entre os modelos avaliados, o GB apresentou maior valor de AUC (0,945), denotando melhor desempenho em distinguir corretamente emulsões estáveis e instáveis. O modelo FA com AUC de 0,925 também apresentou bom desempenho, embora ligeiramente inferior ao GB. Os modelos MLP (0,918) e AD (0,906) obtiveram os menores valores de AUC. Esses modelos podem ser considerados para aplicações de suporte ou em contextos onde os requisitos de desempenho são menos rigorosos.

4.3.1.2 Teste de Friedman para os modelos de classificação

O teste de Friedman e a análise *post-hoc* de Nemenyi foram aplicados para comparar estatisticamente as métricas de desempenho dos modelos de aprendizado de máquina

apresentados na Tabela 3. O estudo foi conduzido com $n = 5$ blocos (métricas: Acurácia, Precisão, Sensibilidade, F1-score, AUC-ROC) e $k = 4$ tratamentos (modelos: AD, GB, FA, MLP), utilizando os valores de desempenho no conjunto de teste. A Tabela 4 apresenta a classificação de cada modelo por métrica e a média geral da classificação.

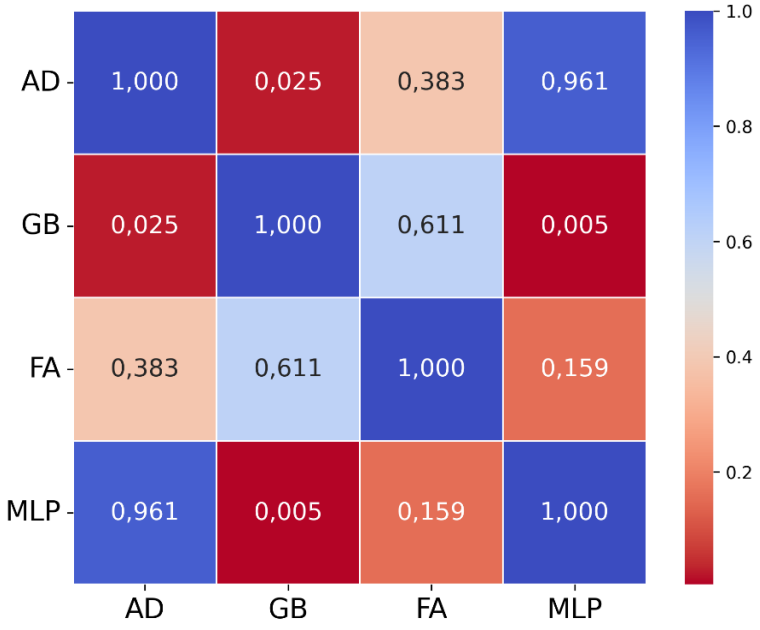
Tabela 4 - Classificação por métrica e média geral.

Métrica	**AD**	**GB**	**FA**	**MLP**
Acurácia	3	1	2	4
Precisão	3	1	2	4
Sensibilidade	3,5	1	2	3,5
F1-Score	3	1	2	4
AUC	4	1	2	3
Média	3,3	1	2	3,7

Fonte: Próprio autor.

O teste de Friedman revelou uma variação significativa entre os modelos ($Q = 14{,}02$, $p = 0{,}003$), rejeitando a hipótese nula e confirmando que ao menos um dos modelos possui desempenho distinto. O teste *post-hoc* de Nemenyi foi realizado para identificar comparações específicas entre os modelos, avaliando se as diferenças pareadas de classificação excedem a Diferença Crítica (DC). A Figura 40 apresenta os *p*-valores do teste de Nemenyi, destacando as diferenças estatisticamente significativas entre os modelos.

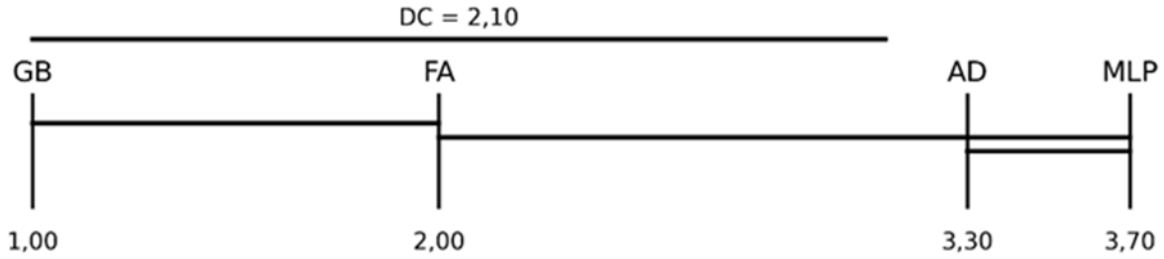
Figura 40 - Matriz de p-valores do teste de Nemenyi para os modelos de estabilidade.



Fonte: Próprio autor.

De acordo com a matriz de *p*-valores, foram observadas diferenças estatisticamente significativas entre GB e AD ($p = 0{,}025$) e entre GB e MLP ($p = 0{,}005$), sugerindo que o modelo *Gradient Boosting* apresenta desempenho significativamente superior a esses modelos em termos das métricas de avaliação no conjunto de teste. A Figura 41 apresenta o diagrama da Diferença Crítica, que facilita a interpretação das comparações pareadas, representando o ranking dos modelos e sua separação estatística.

Figura 41 - Diagrama de diferença crítica para os modelos de classificação.



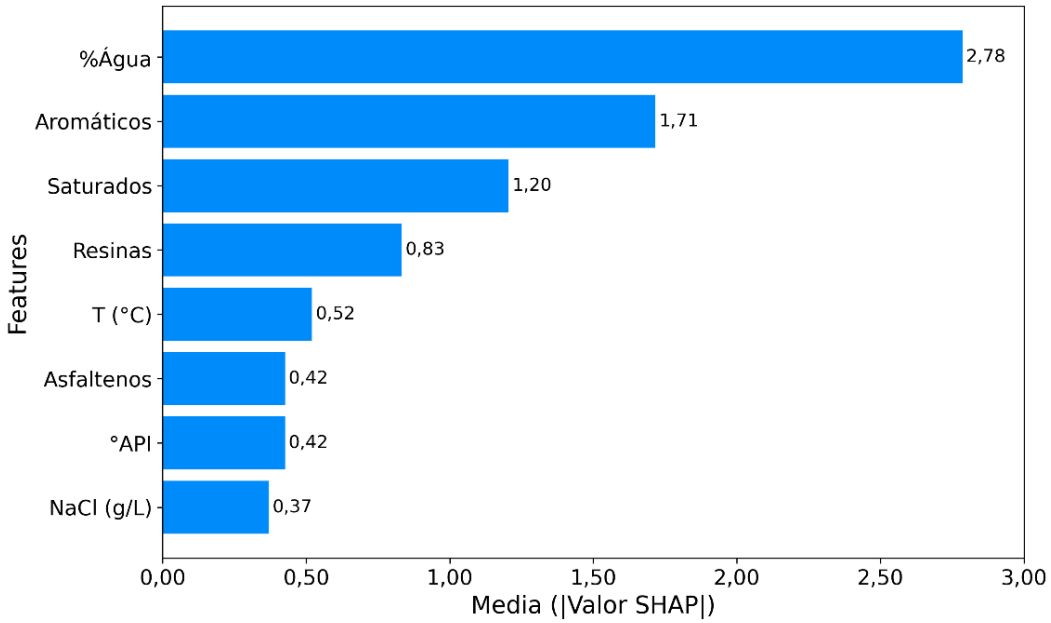
Fonte: Próprio autor.

No diagrama, diferenças nas classificações médias superiores a 2,10 são consideradas estatisticamente significativas. Consequentemente, com uma classificação média de 1, o modelo GB é estatisticamente superior aos modelos DT (3,3) e MLP (3,7). Além disso, os modelos GB e RF apresentam desempenhos estatisticamente equivalentes, indicando que ambos podem ser selecionados como melhor opção.

4.3.1.3 Análise paramétrica com SHAP para a estabilidade de emulsões

Determinar quais parâmetros do conjunto de dados mais influenciam a estabilidade das emulsões fornece informações relevantes. Com base nos resultados das análises estatísticas e de desempenho dos modelos, o modelo *Gradient Boosting* se destacou como o melhor na tarefa de classificação de estabilidade e seus resultados preditivos foram usados para avaliar os parâmetros de entrada. A Figura 42 mostra o valor médio absoluto do impacto que os parâmetros de entrada do modelo tiveram na predição de estabilidade das emulsões.

Figura 42 - Valor médio absoluto dos valores SHAP para predição da estabilidade de emulsões.



Fonte: Próprio autor.

Dos parâmetros de entrada, o teor de água (%Água) foi o mais importante na realização das previsões, seguido dos aromáticos, saturados e resinas. Estudos anteriores relataram que o teor de água desempenha um papel crítico na determinação da eficiência da desestabilização ou da estabilidade de emulsões (Zolfaghari *et al.*, 2016). Além disso, parâmetros como saturados, aromáticos e resinas são fundamentais tanto para a formação quanto para a estabilidade das emulsões, atuando como estabilizadores interfaciais (Romero Yanes *et al.*, 2019; Saad *et al.*, 2019).

De forma a complementar, na Figura 43 apresenta como a magnitude dos parâmetros de entrada do modelo impactaram nas previsões.

Figura 43 - Gráfico de dispersão do impacto das features na saída do modelo para estabilidade de emulsões.

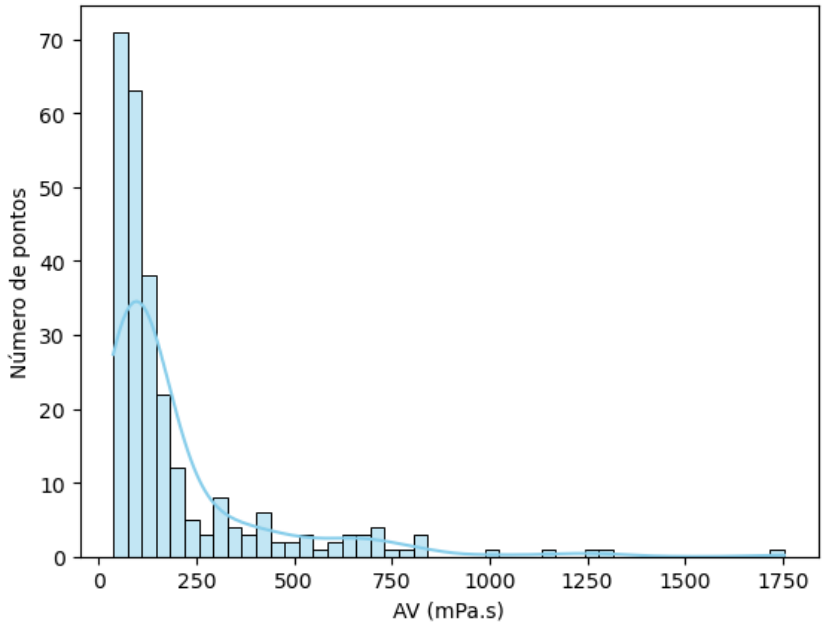
Fonte: Próprio autor.

Dos parâmetros analisados, observa-se que para %Água, maiores teores de água impactaram negativamente na estabilidade de emulsões. Por outro lado, menores teores foram responsáveis pela predição de emulsões estáveis. De acordo com Yonguep *et al.*(2022), Emulsões com maior teor de água tendem a apresentar menor estabilidade, separando-se mais facilmente do que aquelas com menor teor de água.

4.3.2 Viscosidade Aparente

Para realizar predições de viscosidade de emulsões, os modelos *Regressão Polinomial* (RP), *eXtreme Gradient Boosting* (XGBoost), *Floresta Aleatória* (FA) e *Multi-Layer Perceptron* (MLP) foram utilizados. O Conjunto de dados utilizados para a regressão dos dados da viscosidade contém 265 (duzentos e sessenta e cinco) pontos obtidos a partir de 13 (treze) óleos com °API variando de 16,5 até 32,5. A Figura 44 contém a distribuição dos dados em questão.

Figura 44 - Distribuição de dados da viscosidade aparente.



Fonte: Próprio autor.

Para a viscosidade, foi avaliado a eficiência de quatro modelos de aprendizado de máquina (Regressão Polinomial, XGBoost, Random Forest e Redes Neurais Artificiais) para prever realizar a predição de amostras sob diferentes condições experimentais. Esses modelos foram escolhidos devido às suas características distintas, variando desde os modelos mais simples e interpretáveis a métodos mais robustos e não lineares.

A Tabela 5 destaca o desempenho dos modelos ao prever a viscosidade, com cada método oferecendo vantagens específicas.

Tabela 5 - Métricas de aprendizado de máquina de viscosidade para conjunto de treinamento e teste.

Modelos	Métricas							
	R^2		AARE%		RMSD		$\varepsilon_{máx}$ (%)	
	Treino	Teste	Treino	Teste	Treino	Teste	Treino	Teste
RP	0,957	0,953	16,021	16,706	47,858	53,609	121,802	96,985
XGBoost	**0,999**	**0,992**	**2,571**	**10,656**	**8,030**	**22,136**	**16,150**	**30,531**
FA	0,980	0,967	4,875	9,448	32,596	44,786	31,801	35,523
MLP	0,994	0,987	11,405	13,976	17,083	28,043	17,083	52,267

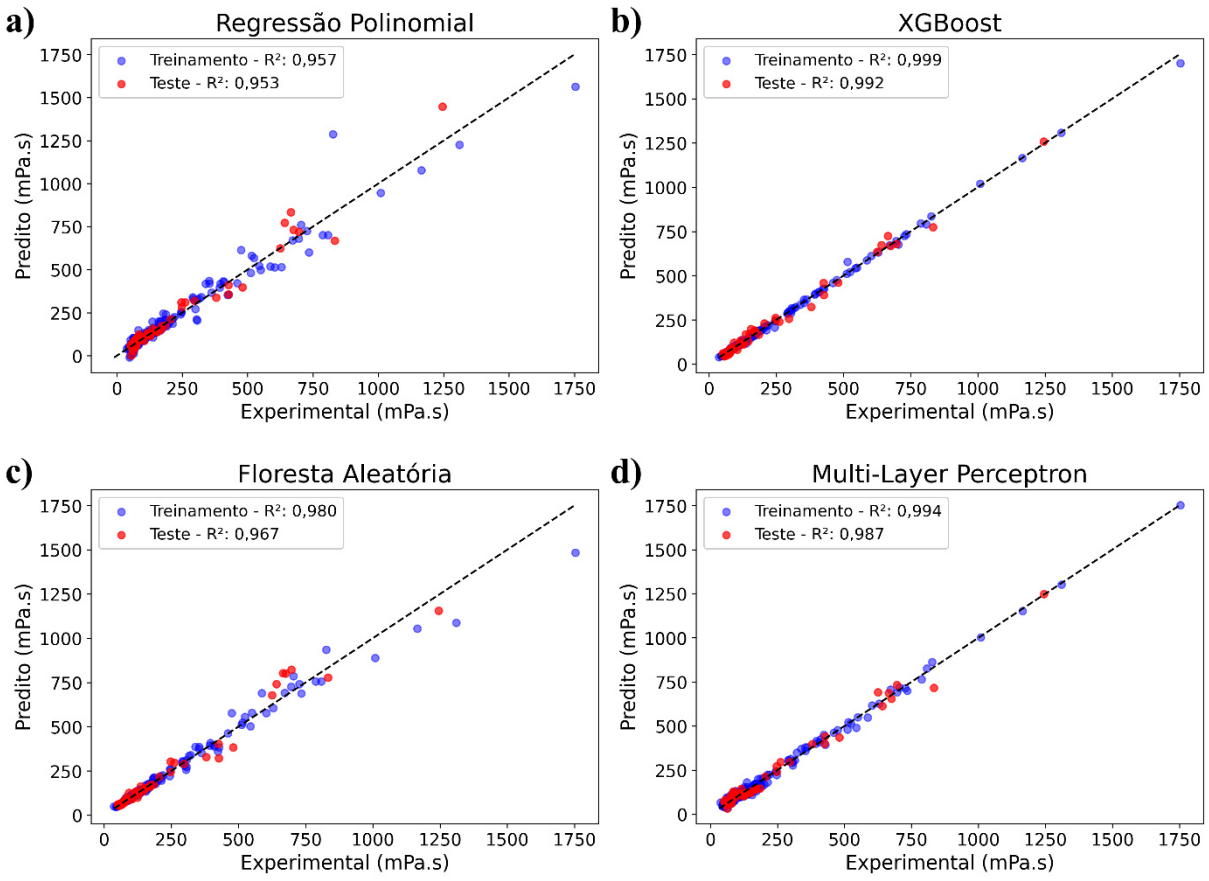
Fonte: Próprio autor.

Entre os modelos avaliados, o XGBoost destacou-se devido à sua alta precisão consistente nos conjuntos de treinamento e teste. Seus valores de R^2 (0,991), AARE% (10,656) e RMSD (22,136) relativamente baixo indicam previsões precisas e uma capacidade de capturar relações complexas nos dados. Além disso, entre os modelos avaliados apresentou o menor erro

absoluto relativo (ε_{max} (%)) no conjunto de teste, sua estabilidade global reforça sua adequação como uma ferramenta confiável para prever viscosidade.

Os modelos FA e MLP obtiveram performances próximas, apresentando forte generalização e desempenho. Os valores de R^2 para esses modelos foram consistentemente acima de 0,96, demonstrando eficiência em explicar a variabilidade dos dados. Apesar do modelo MLP apresentar maior R^2, o modelo FA teve melhor desempenho em relação às métricas AARE% e RMSD durante os testes. A Figura 45 apresenta o gráfico de dispersão para os modelos apresentados na Tabela 5.

Figura 45 - Gráficos de dispersão dos dados de treinamento e teste para a viscosidade aparente.



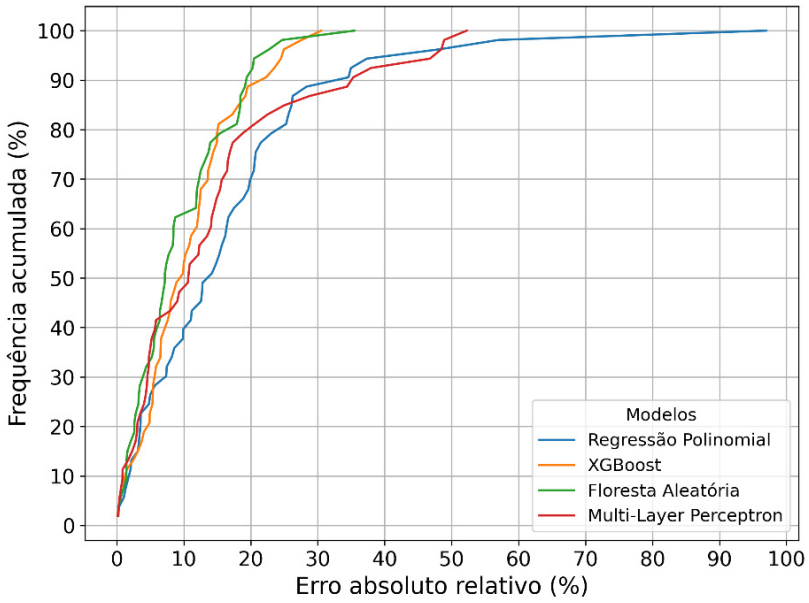
Fonte: Próprio autor.

Em contrapartida, o modelo de Regressão Polinomial (RP) apresentou resultados menos impressionantes. Embora tenha alcançado valores de R^2 razoavelmente altos, seus resultados para AARE% e RMSD foram significativamente maiores, indicando limitações na captura de detalhes intrincados do comportamento da viscosidade. Embora a regressão polinomial possa ser eficaz em condições mais simples, essa comparação ressalta a

superioridade oferecidas por técnicas avançadas de aprendizado de máquina, como XGBoost, FA e MLP.

Para medir o desempenho dos modelos preditivos para a viscosidade de emulsão, foi avaliado como cada modelo se comporta com base no erro absoluto relativo acumulado. A Figura 46 apresenta o comportamento de cada modelo em relação ao erro absoluto relativo.

Figura 46 - Gráfico da frequência acumulada do erro absoluto relativo para predição de viscosidade.



Fonte: Próprio autor.

Os resultados apresentam que os modelos FA e o XGBoost demonstram um crescimento mais rápido na frequência acumulada, atingindo aproximadamente 90% das predições com erro absoluto relativo inferiores a 20%. Isso indica que esses dois modelos são eficientes em prever a viscosidade com eficácia, especialmente nas primeiras faixas de erro.

Por outro lado, o MLP, apesar de também apresentar alta precisão, tem uma curva de crescimento mais gradual, apresentando 90% das predições com erro absoluto relativo entre 30 e 40%. Nesse mesmo intervalo, o modelo RP atinge essa marca antes. Apesar disso, o modelo de Regressão Polinomial, é o modelo menos eficiente nesse conjunto, apresentando o maior $\varepsilon_{máx}$ e atingindo 100% da frequência acumulada por último.

4.3.2.1 Teste de Friedman para os modelos de viscosidade aparente

Para comparar estatisticamente o desempenho dos modelos de predição de viscosidade a partir dos resultados apresentados na Tabela 5, aplicou-se o teste de Friedman seguido da

análise *post-hoc* de Nemenyi. A análise considerou os valores de desempenho obtidos no conjunto de teste, com n = 4 blocos (métricas: R^2, AARE%, RMSD, $\varepsilon_{máx}$ (%)) e k = 4 tratamentos (modelos: RP, XGBoost, FA, MLP). A Tabela 6 exibe os postos de cada modelo por métrica e a média geral dos postos.

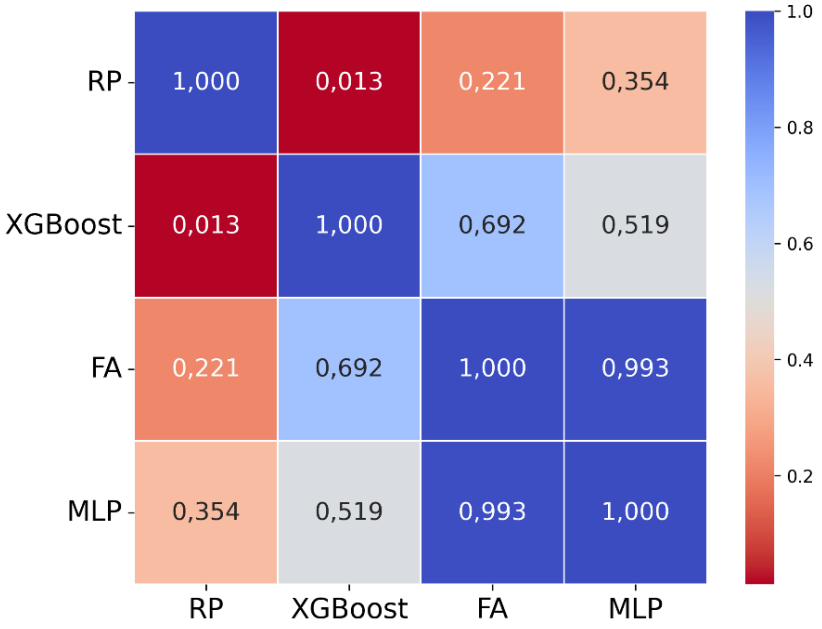
Tabela 6 - Classificação por métrica e média geral.

Métrica	RP	XGBoost	FA	MLP
R^2	4	1	3	2
AARE%	4	2	1	3
RMSD	4	1	3	2
$\varepsilon_{máx}$ (%)	4	1	2	3
Média	4	1,25	2,25	3

Fonte: Próprio autor.

O teste de Friedman revelou uma variação significativa entre os modelos (Q = 9,30, *p* = 0,0256), rejeitando a hipótese nula e confirmando que ao menos um dos modelos possui desempenho distinto. O teste *post-hoc* de Nemenyi foi realizado e os *p*-valores dispostos na Figura 47.

Figura 47 - matriz de *p*-valores do teste de Nemenyi.

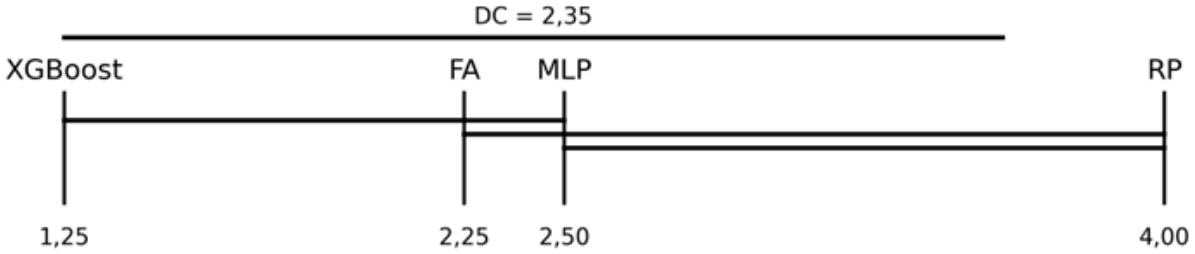


Fonte: Próprio autor.

De acordo com a matriz de *p*-valores, foram observadas diferenças estatisticamente significativas entre XGBoost e RP (*p* = 0,013), sugerindo que o modelo XGBoost apresenta

desempenho significativamente superior em termos das métricas usadas para avaliação dos modelos no conjunto de teste. A Figura 48 apresenta o diagrama da diferença crítica, que facilita a interpretação das comparações pareadas, representando a classificação dos modelos e sua separação estatística.

Figura 48 - Diagrama de diferença crítica para os modelos de viscosidade.



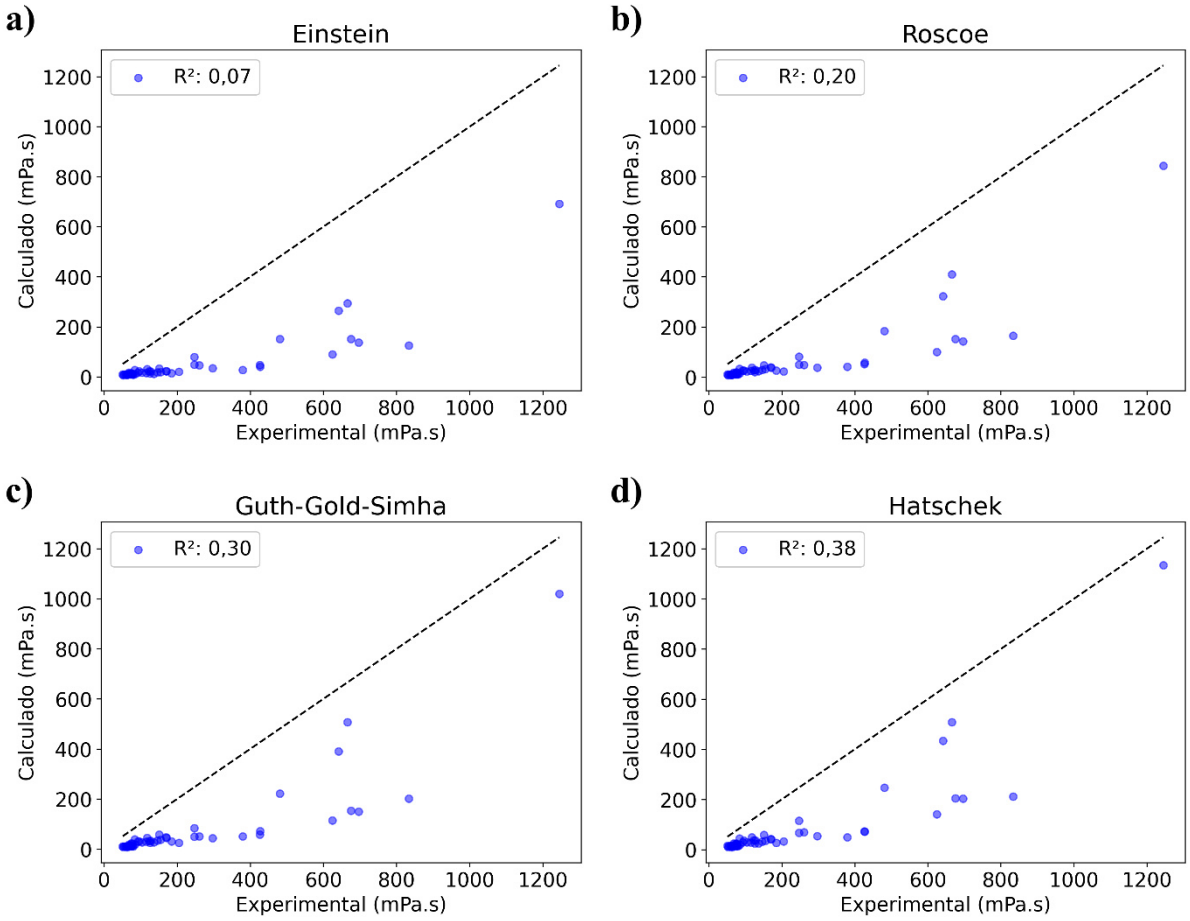
Fonte: Próprio autor.

No diagrama da diferença crítica, diferenças nas classificações médias superiores a 2,35 são consideradas estatisticamente significativas. Consequentemente, com uma classificação média de 1,25, o modelo XGBoost é estatisticamente superior ao modelo RP (4). Assim, XGBoost, FA e MLP apresentam desempenhos estatisticamente equivalentes, indicando que todos podem ser selecionados como melhor opção.

4.3.2.2 Comparativo entre correlações empíricas clássicas

As correlações empíricas, como as de Einstein, Roscoe, Guth-Gold-Simha e Hatschek, representam abordagens clássicas baseadas em relações matemáticas predefinidas para descrever o comportamento da viscosidade em função da fração da fase dispersa. Elas oferecem a opção para estimativas devido a simplicidade. No entanto, essas correlações apresentam limitações, especialmente quando aplicadas a sistemas complexos ou dados que fogem dos pressupostos básicos dessas fórmulas. A Figura 49 é apresentado o desempenho das 4 (quatro) correlações empíricas clássicas citadas anteriormente quando testadas para o mesmo corte de água das predições realizadas pelos métodos de AM.

Figura 49 - Comparativo dos valores de viscosidade calculados a partir das correlações empíricas clássicas na literatura e os valores experimentais.



Fonte: Próprio autor.

Ao analisar os coeficientes de determinação R^2 de cada correlação, fica evidente que, para o problema em questão que envolve múltiplas variáveis que afetam a viscosidade das emulsões, elas não alcançam bons níveis de precisão, visto que elas estão limitadas somente à fração de água na emulsão. Isso se deve, em parte, ao fato de que modelos empíricos não possuem termos que levem em conta outros fatores que afetam a viscosidade além da fração de água dispersa na emulsão. Comparativamente, o modelo de Regressão Polinomial que obteve o menor desempenho alcançou R^2 de 0,953 no conjunto de teste, demonstrou uma capacidade significativamente superior em ajustar os dados experimentais.

A vantagem de modelos de inteligência artificial reside na sua flexibilidade e capacidade de aprendizado com base em grandes conjuntos de dados. Diferentemente das correlações empíricas, o modelo de IA não está limitado a uma estrutura matemática específica; ele é capaz de identificar padrões e relações sutis nos dados, ajustando suas predições de forma

dinâmica. Isso resulta em uma precisão muito maior, especialmente em casos onde variáveis adicionais ou comportamento não-linear estão presentes.

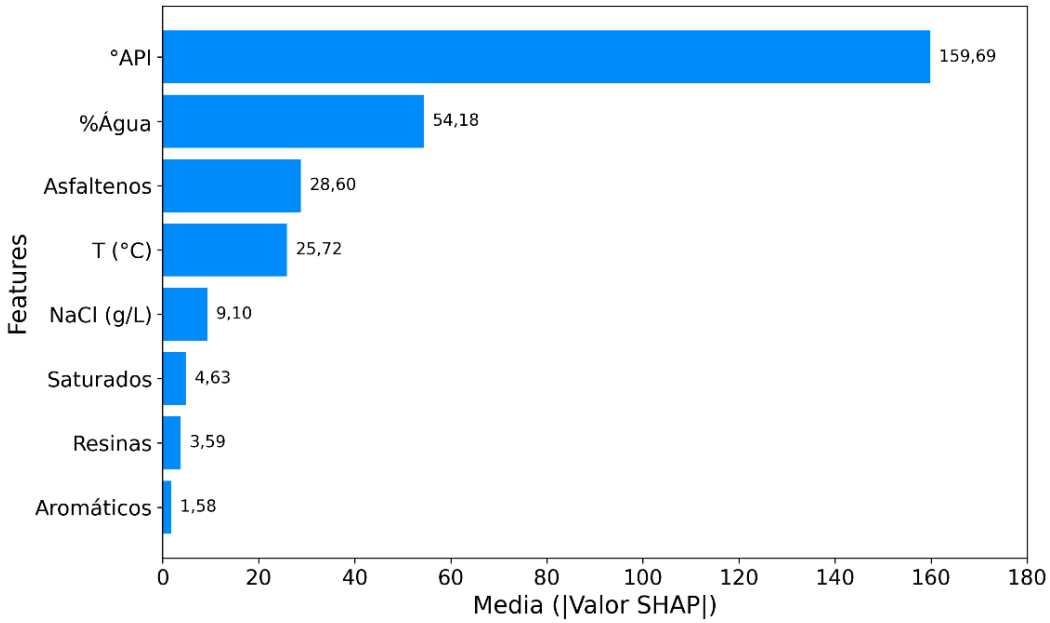
4.3.2.3 Análise paramétrica com SHAP para a viscosidade

Com base nas métricas de avaliação, na análise gráfica e nos resultados do teste de Friedman, o modelo XGBoost foi identificado como o de melhor desempenho. Com isso, as predições deste modelo foram submetidas à análise SHAP (*SHapley Additive exPlanations*) para examinar a influência das variáveis de entrada nas previsões de viscosidade.

O valor SHAP médio absoluto foi utilizado para representar a importância relativa de cada variável de entrada. Este valor, calculado pela média dos valores SHAP absolutos para cada variável, reflete seu impacto médio na saída do modelo, fornecendo uma medida clara das variáveis mais influentes e auxiliando na interpretação do comportamento do modelo.

A Figura 50 ilustra a importância das variáveis para a predição da viscosidade de emulsões, destacando a gravidade API, a %Água, os Asfaltenos e a Temperatua como os parâmetros de entrada mais influentes, classificados nessa ordem.

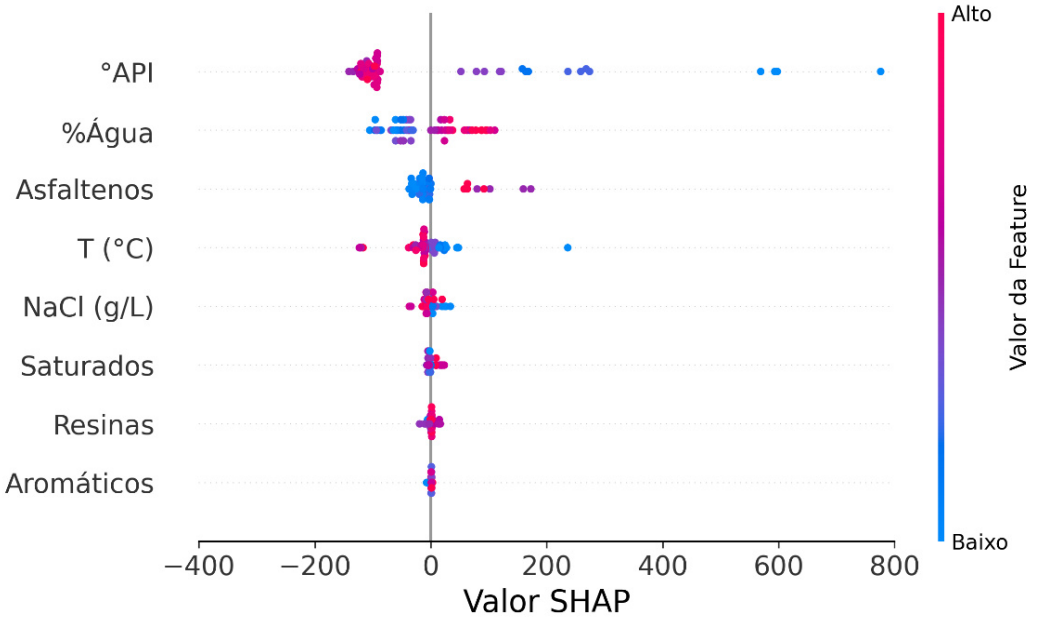
Figura 50 - Valor médio absoluto dos valores SHAP para predição de viscosidade de emulsões.



Fonte: Próprio autor.

Complementarmente as *features* que representam maior importância para previsão do modelo, o gráfico de dispersão ajuda a entender como os valores de cada feature impactam na saída do modelo. A influência de cada uma delas pode ser observada na Figura 51.

Figura 51 - Gráfico de dispersão do impacto das features na saída do modelo para viscosidade.



Fonte: Próprio autor.

Entre os parâmetros de entrada, o °API exerce maior influência, reduzindo a viscosidade em valores altos e aumentando em valores baixos, devendo-se à correlação inversa entre densidade e fluidez do petróleo. A %Água afeta diretamente a interação das fases e contribui para variações na viscosidade evidenciando-se pela indicação que valores mais altos retornam valores mais altos de viscosidade pelo modelo, enquanto valores menores faz o modelo retornar viscosidades menores.

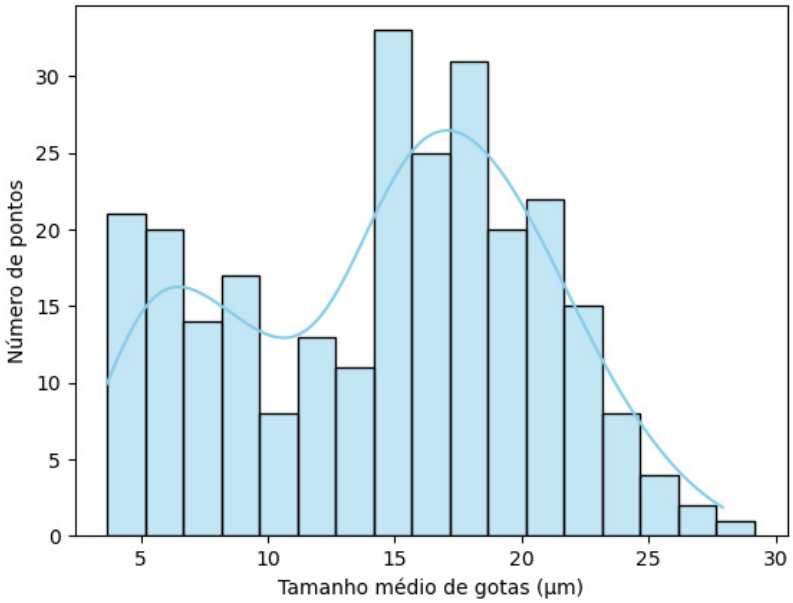
Os asfaltenos e a temperatura também têm impacto relevante, com valores maiores de asfaltenos elevando a viscosidade e temperaturas mais altas promovendo redução. Concentração de sal, saturados, resinas e aromáticos foram menos influentes nas predições realizadas pelo modelo.

4.3.3 Diâmetro Médio de Gota

Assim como para a viscosidade, para realizar predições de viscosidade de emulsões, os modelos *Regressão Polinomial* (RP), *eXtreme Gradient Boosting* (XGBoost), *Floresta*

Aleatória (FA) e *Multi-Layer Perceptron* (MLP) foram utilizados. O conjunto de dados utilizados para a regressão dos dados de diâmetro médio de gota (DMG) contém 265 (duzentos e sessenta e cinco) pontos obtidos a partir de 13 (treze) óleos com °API variando de 16,5 até 32,5. A Figura 52 contém a distribuição dos dados em questão.

Figura 52 - Distribuição dos tamanhos médios de gotas



Fonte: Próprio autor.

Os modelos de aprendizado de máquina investigados apresentaram desempenhos variados na previsão de DMG em emulsões água-em-óleo (A/O), exibindo diferentes níveis de precisão e capacidade de capturar padrões nos dados experimentais. O desempenho de cada um dos modelos está apresentado na Tabela 7, que fornece uma comparação detalhada entre o desempenho dos modelos nas etapas de treinamento e teste.

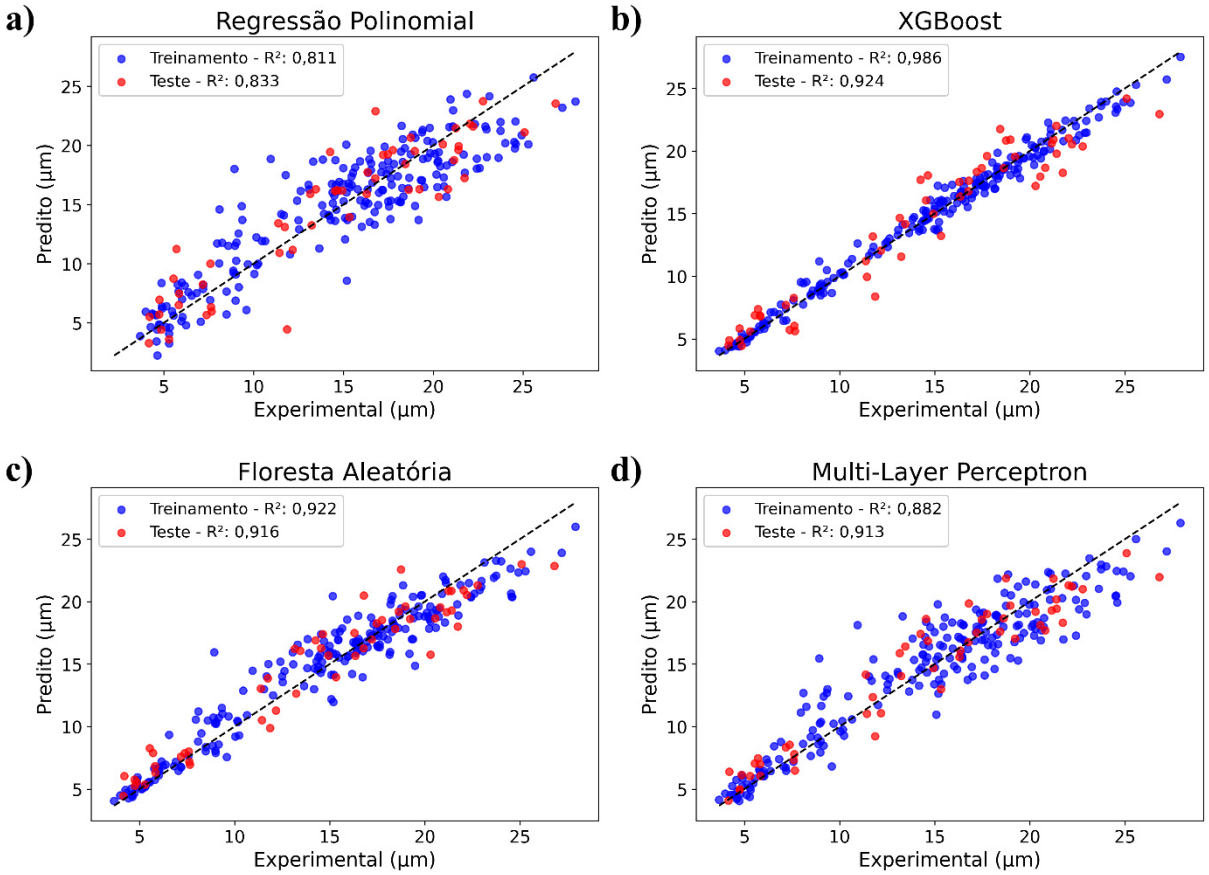
Tabela 7 - Métricas de aprendizado de máquina de DMG para conjunto de treinamento e teste.

Modelos	Métricas							
	R^2		AARE%		RMSD		$\varepsilon_{máx}$ (%)	
	Treino	Teste	Treino	Teste	Treino	Teste	Treino	Teste
RP	0,811	0,833	16,130	17,760	2,564	2,581	102,071	97,151
XGBoost	**0,986**	**0,924**	**3,753**	**11,025**	**0,695**	**1,738**	**25,485**	**29,284**
FA	0,922	0,916	8,996	11,733	1,651	1,829	78,815	49,171
MLP	0,882	0,913	11,286	11,811	2,030	1,866	73,294	52,644

Fonte: Próprio autor.

O desempenho dos modelos na previsão do diâmetro médio de gota destaca diferenças notáveis em suas capacidades preditivas. Entre eles, o modelo XGBoost apresentou melhor desempenho entre todas as métricas avaliadas, apresentando os maiores valores de R^2 para o conjunto de dados de teste (0,924), além dos menores valores de AARE% (11,025), RMSD (1,738) e $\varepsilon_{máx}$ (%) (29,284). A Figura 53 apresenta os gráficos de dispersão para os modelos apresentados na Tabela 7.

Figura 53 - Gráfico de correlação dos dados de treinamento e teste para o tamanho médio das gotas.



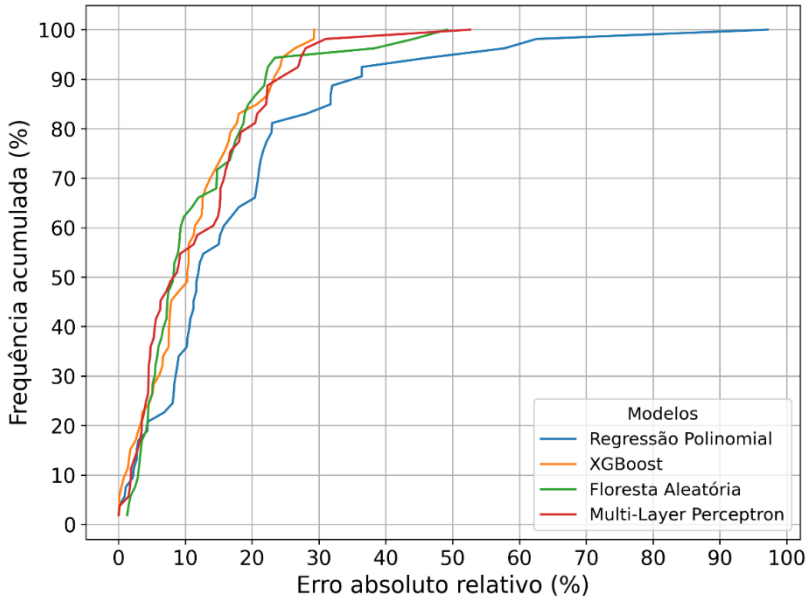
Fonte: Próprio autor.

Os modelos FA e MLP apresentaram desempenhos próximos ao XGBoost, ambos alcançaram valores relativamente altos de R^2 e valores baixos AARE% (11,733 e 11,811) e RMSD (1,829 e 1,866), embora o FA tenha exibido uma consistência ligeiramente melhor entre os conjuntos de dados de treinamento e teste. Por outro lado, o MLP, apesar de sua precisão ligeiramente inferior, mostrou fortes capacidades em capturar relações não lineares, tornando-se uma alternativa viável dependendo do contexto e das exigências da análise. Contudo, ambos apresentaram valores relativamente altos para $\varepsilon_{máx}$ (%), com o FA apresentando valor 49,171 e MLP atingindo 52,644.

Em contraste, o modelo RP apresentou o pior desempenho em todas as métricas avaliadas, com valores de R^2 inferiores e AARE%, RMSD e $\varepsilon_{máx}$ (%) mais elevados em relação aos outros modelos. Esse resultado provavelmente reflete as limitações da regressão polinomial em capturar a complexidade e as interações não lineares do conjunto de dados. Embora a RP ainda possa fornecer informações relevantes em condições mais simples, sua precisão comparativamente limitada destaca as vantagens de utilizar métodos mais avançados (como XGBoost, FA e MLP) para tarefas preditivas complexas envolvendo o DMG.

Em contrapartida o desempenho dos modelos preditivos para a o diâmetro médio de gota foi avaliado em como cada modelo se comporta com base no erro absoluto relativo acumulado, assim como para viscosidade. A Figura 54 apresenta o comportamento de cada modelo em relação ao erro absoluto relativo.

Figura 54 - Gráfico da frequência acumulada do erro absoluto relativo para predição de DMG.



Fonte: Próprio autor.

Os resultados apresentam que os modelos FA e o XGBoost demonstram um crescimento mais rápido na frequência acumulada, atingindo aproximadamente 90% das predições com erro absoluto relativo inferiores a 30%, com uma ligeira vantagem do modelo FA. Contudo, o modelo XGBoost atinge 100% das predições com erro absoluto relativo inferior a 30%, enquanto o modelo FA atinge a mesma marca entre 40 e 50%, evidenciando a superioridade do XGBoost.

Por outro lado, o MLP também apresenta alta precisão em relação aos modelos anteriores, atingindo 90% das predições com erro absoluto relativo também menor que 30%.

ficando atrás do modelo RP que atinge essa marca antes. Apesar disso, o modelo de Regressão Polinomial, é o modelo menos eficiente nesse conjunto, apresentando o maior $\varepsilon_{máx}$ e atingindo 100% da frequência acumulada por último.

A Regressão Polinomial, por sua vez, exibe limitações claras. É significativamente mais lenta e imprecisa, atingindo 90% das predições com erro absoluto relativo entre 30 e 40%. Além disso, 100% das predições acumuladas é tingido quando o erro absoluto relativo está próximo de 100%, indicando que o modelo RP apresenta grandes erros nas predições, evidenciando seu desempenho inferior.

4.3.3.1 Teste de Friedman para os modelos de diâmetro médio de gota

Para comparar estatisticamente o desempenho dos modelos de predição de DMG a partir dos resultados apresentados na Tabela 7, aplicou-se o teste de Friedman seguido da análise *post-hoc* de Nemenyi. A análise considerou os valores de desempenho obtidos no conjunto de teste, com $n = 4$ blocos (métricas: R^2, AARE%, RMSD, $\varepsilon_{máx}$ (%)) e $k = 4$ tratamentos (modelos: RP, XGBoost, FA, MLP). A Tabela 8 exibe os postos de cada modelo por métrica e a média geral dos postos.

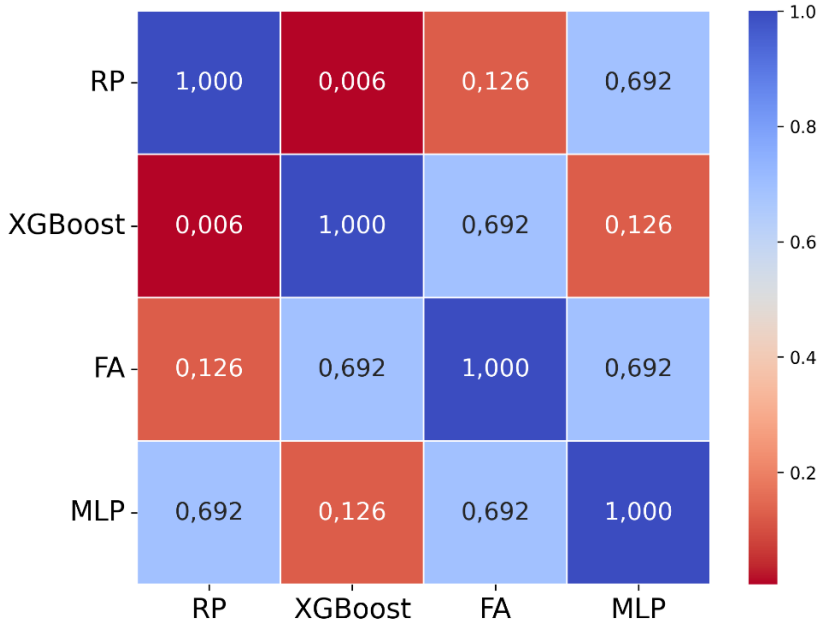
Tabela 8 - Classificação por métrica e média geral.

Métrica	**RP**	**XGBoost**	**FA**	**MLP**
R^2	4	1	2	3
AARE%	4	1	2	3
RMSD	4	1	2	3
$\varepsilon_{máx}$ (%)	4	1	2	3
Média	4	1	2	3

Fonte: Próprio autor.

O teste de Friedman revelou uma variação significativa entre os modelos ($Q = 12{,}00$, $p = 0{,}0074$), rejeitando a hipótese nula e confirmando que ao menos um dos modelos possui desempenho distinto. Um teste *post-hoc* de Nemenyi foi realizado para identificar comparações específicas entre os modelos, avaliando se as diferenças pareadas de ranking excedem a Diferença Crítica (DC). A Figura 55 apresenta os *p*-valores do teste de Nemenyi, destacando as diferenças estatisticamente significativas entre os modelos.

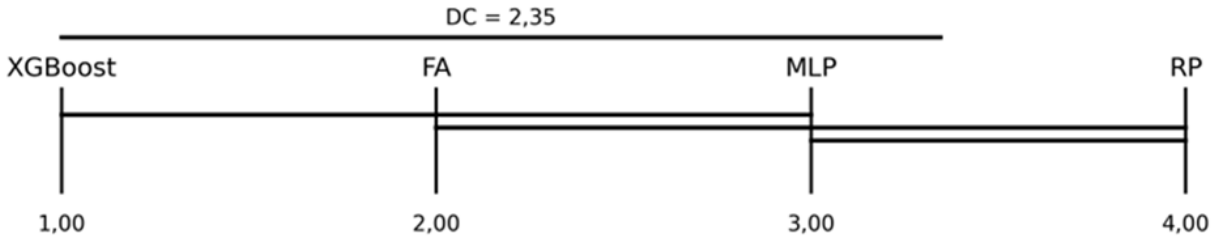
Figura 55 - Matriz de *p*-valores do teste de Nemenyi.



Fonte: Próprio autor.

De acordo com a matriz de *p*-valores, foram observadas diferenças estatisticamente significativas entre XGBoost e RP (p = 0,006), sugerindo que o modelo XGBoost apresenta desempenho significativamente superior em termos das métricas usadas para avaliação dos modelos no conjunto de teste. A Figura 56 apresenta o diagrama da diferença crítica, que facilita a interpretação das comparações pareadas, representando a classificação dos modelos e sua separação estatística.

Figura 56 - Diagrama de diferença crítica para os modelos de DMG.



Fonte: Próprio autor.

No diagrama, diferenças nas classificações médias superiores a 2,35 são consideradas estatisticamente significativas. Consequentemente, com uma classificação média de 1, o modelo XGBoost é estatisticamente superior ao modelo RP (4). Assim, XGBoost, FA e MLP

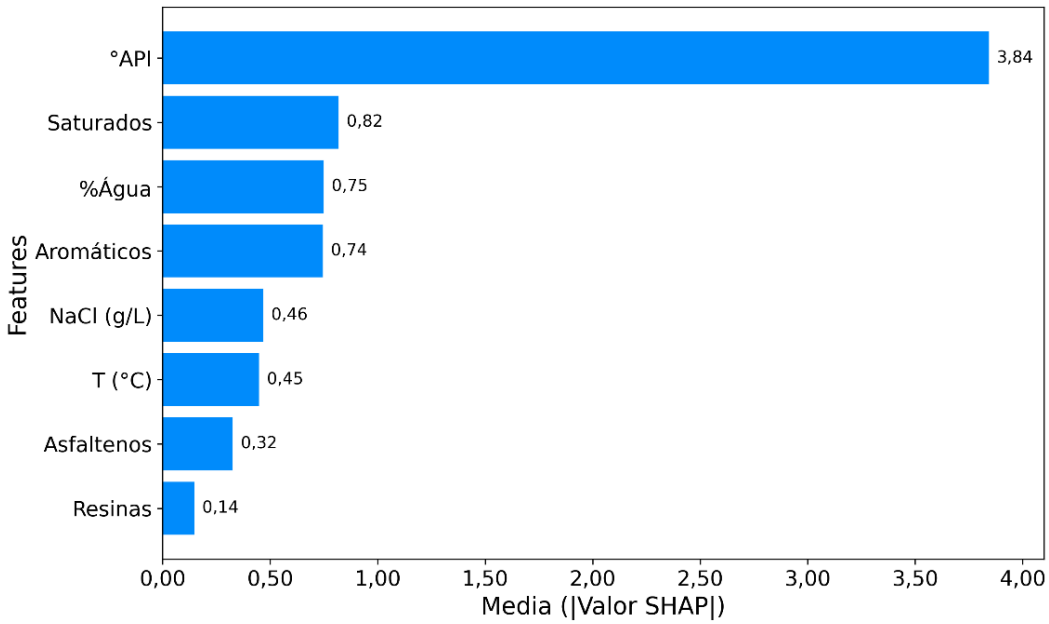
apresentam desempenhos estatisticamente equivalentes, indicando que todos modelos pode ser selecionado como melhor opção.

4.3.3.2 Análise paramétrica com SHAP para DMG

Com base nas métricas de avaliação, na análise gráfica e nos resultados do teste de Friedman, o modelo XGBoost foi identificado como o de melhor desempenho. Com isso, as predições deste modelo foram submetidas à análise SHAP para examinar a influência das variáveis de entrada nas previsões de viscosidade.

A Figura 57 ilustra a importância das variáveis para a determinação do diâmetro médio de gota. Nessa ordem, o °API, teor de água (%Água), saturados, temperatura (°C) e concentração de sal na salmoura (NaCl) são os parâmetros de entrada mais significativos.

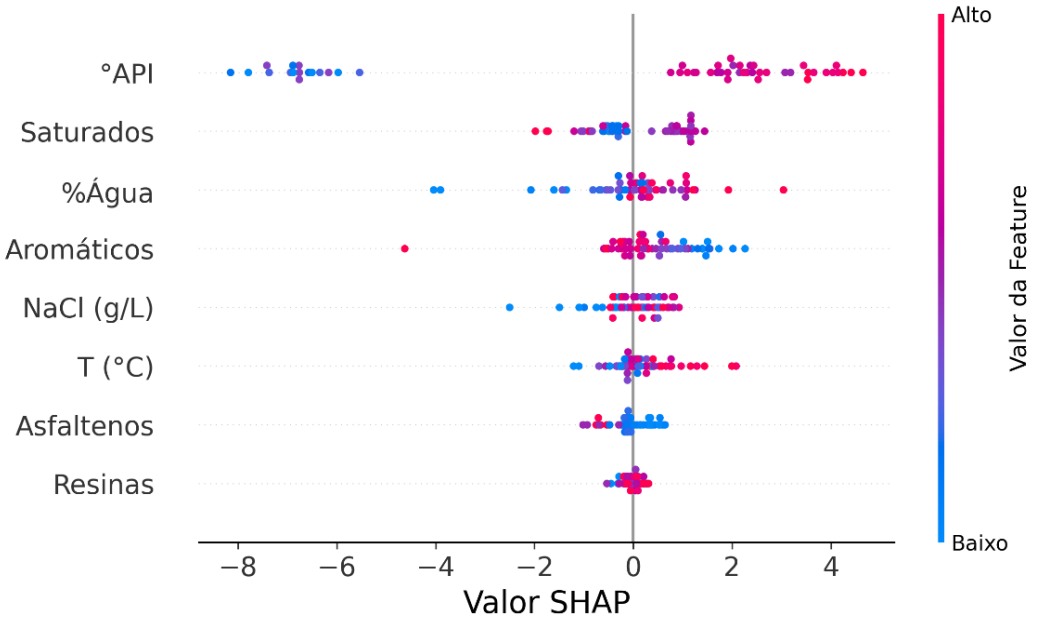
Figura 57 - Valor médio absoluto dos valores SHAP para predição de diâmetro médio de gota em emulsões.



Fonte: Próprio autor.

Complementarmente, a Figura 58 indica como os valores de cada feature impactam nos valores preditos na saída do modelo.

Figura 58 - Gráfico de dispersão do impacto das features na saída do modelo para DMG.



Fonte: Próprio autor.

Assim como para a viscosidade o °API ainda é a variável mais influente, mostrando que valores mais altos estão correlacionados com um aumento no tamanho das gotas, provavelmente devido a facilidade com que as gotas de água têm de se movimentarem e consequentemente promover mais choques entre as gotas, favorecendo a coalescência e a formação de gotas maiores. Já para valores menores de °API, as gotas de água tem sua mobilidade reduzida, dificultando a coalescência (Ferreira Filho *et al.*, 2020). A %Água desempenha um papel essencial, já que concentrações maiores de água na emulsão facilitam interações mais frequentes entre as gotas, contribuindo diretamente para o aumento do tamanho.

Como observado na importância das variáveis SHAP, o que está relatado na literatura, a gravidade API e o teor de água influenciam a distribuição do tamanho das gotas. À medida que a concentração da fase dispersa (água) aumenta, a reologia da dispersão muda significativamente devido ao aumento da frequência de interação entre as gotas (Ilia Anisa; Nour, 2010; Liu *et al.*, 2024).

5 CONCLUSÃO

No presente trabalho, o uso do método de Amostragem por Hipercubo Latino (LHS), permitiu a seleção aleatória de valores de temperatura, concentração de NaCl e de corte de água para testar experimentalmente a influência delas sobre um conjunto de diferentes óleos oriundos de 13 (treze) diferentes poços.

Utilizaram-se como dados de entrada para os modelos de aprendizado de máquina os valores experimentais gerados pelo método LHS, além de propriedades como °API e a composição SARA para diferenciação das diferentes amostras de óleo para predição de estabilidade, viscosidade e diâmetro médio de gota de sistemas emulsionados A/O.

Foi desenvolvido um algoritmo, baseado na Transformada de Hough para a detecção automática das gotas de água presentes nas imagens das emulsões, com o uso da biblioteca OpenCV em Python. Essa abordagem permitiu identificar padrões circulares em diferentes condições experimentais, permitindo maior velocidade e precisão na quantificação do diâmetro médio das gotas.

Os modelos treinados e avaliados que apresentaram melhor desempenho para cada propriedade avaliada foram baseados em técnicas de *Gradient Boosting*. O modelo GB se destacou na previsão da estabilidade, enquanto o XGBoost foi mais eficaz na estimativa da viscosidade e do diâmetro. Para predição de estabilidade de emulsões, o modelo GB performou acima dos demais modelos em todas as métricas avaliadas (Acurácia = 0,981; Precisão = 0,974; Sensibilidade = 0,925; F1-Score = 0,949; AUC-ROC = 0,945). Para a viscosidade, o modelo XGBoost obteve melhores resultados em 3 (três) das 4 métricas avaliadas (R^2 = 0,992; RMSD = 22,136; $\varepsilon_{máx}$ (%) = 30,531). Por fim, para o diâmetro médio de gota, o modelo XGBoost atingiu melhores marcas nas 4 métricas avaliadas (R^2 = 0,924; AARE% = 11,025; RMSD = 1,738; $\varepsilon_{máx}$ (%) = 29,284.

Esse desempenho superior pode estar relacionado ao funcionamento dos métodos de boosting, que constroem os modelos de forma sequencial e aditiva, permitindo que cada nova árvore corrija os erros cometidos pelas anteriores. Isso pode ter contribuído para maior precisão e capacidade de generalização dos modelos.

REFERÊNCIAS

ABDEL-RAOUF, Manar El-Sayed. **Crude Oil Emulsions- Composition Stability and Characterization**. [*S. l.*]: InTech, 2012. Disponível em: http://www.intechopen.com/books/crude-oil-emulsions-composition-stability-and-characterization.

ABDULREDHA, Murtada Mohammed; SITI ASLINA, Hussain; LUQMAN, Chuah Abdullah. Overview on petroleum emulsions, formation, influence and demulsification treatment techniques. **Arabian Journal of Chemistry**, [*s. l.*], v. 13, n. 1, p. 3403–3428, 2020. Disponível em: https://linkinghub.elsevier.com/retrieve/pii/S1878535218302442.

AHMETOGLU, Huseyin; DAS, Resul. A comprehensive review on detection of cyber-attacks: Data sets, methods, challenges, and future research directions. **Internet of Things**, [*s. l.*], v. 20, p. 100615, 2022. Disponível em: https://linkinghub.elsevier.com/retrieve/pii/S254266052200097X.

AÏFA, Tahar. Neural network applications to reservoirs: Physics-based models and data models. **Journal of Petroleum Science and Engineering**, [*s. l.*], v. 123, p. 1–6, 2014. Disponível em: https://linkinghub.elsevier.com/retrieve/pii/S0920410514003441.

AL-MARHOUN, Muhammad Ali. A Single Artificial Neural Network Model Predicts Bubble Point Physical Properties of Crude Oils. *In*: , 2021. **SPE Middle East Oil and Gas Show and Conference, MEOS, Proceedings**. [*S. l.*]: SPE, 2021. Disponível em: https://onepetro.org/SPEMEOS/proceedings/21MEOS/21MEOS/D021S006R006/474466.

ALPAYDIN, Ethem. **Introduction to Machine Learning**. [*S. l.: s. n.*], 2020. Disponível em: https://mitpress.mit.edu/9780262043793/introduction-to-machine-learning/.

ALVES, Raissa S. **Avaliação de aditivos baseados em óleo de mamona como desemulsificantes de emulsões tipo água em óleo**. 2020. 101 f. - Universidade Federal do Ceará (UFC), [*s. l.*], 2020. Disponível em: https://repositorio.ufc.br/bitstream/riufc/51325/5/2020_dis_rsalves.pdf.

AMIRIAN, Ehsan *et al.* Integrated cluster analysis and artificial neural network modeling for steam-assisted gravity drainage performance prediction in heterogeneous reservoirs. **Expert Systems with Applications**, [*s. l.*], v. 42, n. 2, p. 723–740, 2015. Disponível em: https://linkinghub.elsevier.com/retrieve/pii/S0957417414005168.

ANCHEYTA, Jorge. **Modeling of Processes and Reactors for Upgrading of Heavy Petroleum**. [*S. l.: s. n.*], 2013.

ANDRADE, Lucas H. S. de; COSTA, Bruno L. G.; ANGÉLICO, Bruno A. PSO aplicado à sintonia do controlador Pi/Pid da malha de nível de uma planta didática industrial. **XI SIMPÓSIO BRASILEIRO DE AUTOMAÇÃO INTELIGENTE**, [*s. l.*], p. 1–6, 2013. Disponível em: http://www.sbai2013.ufc.br/pdfs/7935.pdf.

ANP. **Boletim da Produção de Petróleo e Gás Natural - 2024**. [*S. l.: s. n.*], 2024.

ARJARIA, Siddhartha Kumar; RATHORE, Abhishek Singh; CHERIAN, Jincy S. Kidney disease prediction using a machine learning approach: A comparative and comprehensive analysis. *In*: DEMYSTIFYING BIG DATA, MACHINE LEARNING, AND DEEP LEARNING FOR HEALTHCARE ANALYTICS. [*S. l.*]: Elsevier, 2021. p. 307–333. Disponível em: https://linkinghub.elsevier.com/retrieve/pii/B9780128216330000064.

AZODI, Masood; SOLAIMANY NAZAR, Ali Reza. An experimental study on factors affecting the heavy crude oil in water emulsions viscosity. **Journal of Petroleum Science and Engineering**, [*s. l.*], v. 106, p. 1–8, 2013. Disponível em: https://linkinghub.elsevier.com/retrieve/pii/S0920410513001034.

BAKRY, Amr M. *et al.* Microencapsulation of Oils: A Comprehensive Review of Benefits, Techniques, and Applications. **Comprehensive Reviews in Food Science and Food Safety**, [*s. l.*], v. 15, n. 1, p. 143–182, 2016. Disponível em: https://ift.onlinelibrary.wiley.com/doi/10.1111/1541-4337.12179.

BANGERT, Patrick. **Machine Learning and Data Science in the Oil and Gas Industry: Best Practices, Tools, and Case Studies**. [*S. l.*]: Elsevier, 2021. Disponível em: https://linkinghub.elsevier.com/retrieve/pii/C2019002033X.

BELYADI, Hoss; HAGHIGHAT, Alireza. **Machine Learning Guide for Oil and Gas Using Python: A Step-by-Step Breakdown with Data, Algorithms, Codes, and Applications**. [*S. l.*]: Elsevier, 2021. Disponível em: https://linkinghub.elsevier.com/retrieve/pii/C20190036175.

BORGES, Patrícia Carolina Santos. **Otimização Dinâmica Da Fermentação Alcoólica No Processo Em Batelada Alimentada**. 2008. 162 f. - Universidade Federal de Uberlândia (UFU), [*s. l.*], 2008. Disponível em: https://repositorio.ufu.br/bitstream/123456789/15105/1/Patricia.pdf.

BRADLEY, Andrew P. The use of the area under the ROC curve in the evaluation of machine learning algorithms. **Pattern Recognition**, [*s. l.*], v. 30, n. 7, p. 1145–1159, 1997. Disponível em: https://linkinghub.elsevier.com/retrieve/pii/S0031320396001422.

BURKE, Inga *et al.* Two deep learning methods in comparison to characterize droplet sizes in emulsification flow processes. **Journal of Flow Chemistry**, [*s. l.*], v. 14, n. 4, p. 597–613, 2024. Disponível em: https://link.springer.com/10.1007/s41981-024-00330-3.

CARREÓN, Bernardo *et al.* **Petroleum Engineering Thermophysical Properties of Heavy Petroleum Fluids**. Cham: Springer International Publishing, 2021-. ISSN 2366-2646.(Petroleum Engineering). Disponível em: http://www.springer.com/series/15095.

CHEN, Tianqi; GUESTRIN, Carlos. XGBoost: A scalable tree boosting system. *In*: , 2016, New York, NY, USA. **Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining**. New York, NY, USA: ACM, 2016. p. 785–794. Disponível em: https://dl.acm.org/doi/10.1145/2939672.2939785.

CHOLLET, François. Deep Learning with Python, Second Edition. **Deep Learning with Python**, [*s. l.*], 2021. Disponível em: https://www.manning.com/books/deep-learning-with-python-second-edition.

COELHO, Leandro dos Santos; KROHLING, Renato A. Controlador Preditivo Baseado em Otimização por Colônia de Partículas. *In*: , 2003. **Anais do 6. Congresso Brasileiro de Redes Neurais**. [*S. l.*]: SBRN, 2003. p. 259–264. Disponível em: http://abricom.org.br/eventos/cbrn_2003/6CBRN_056.

CUNHA, Roberto E.P. *et al.* Mathematical modeling of the destabilization of crude oil emulsions using population balance equation. **Industrial and Engineering Chemistry Research**, [*s. l.*], v. 47, n. 18, p. 7094–7103, 2008. Disponível em: https://pubs.acs.org/doi/10.1021/ie800391v.

DE OLIVEIRA, Cesar B.Z. *et al.* Rheological Properties of Water-in-Brazilian Crude Oil Emulsions: Effect of Water Content, Salinity, and pH. **Energy and Fuels**, [*s. l.*], v. 32, n. 8, p. 8880–8890, 2018. Disponível em: https://pubs.acs.org/doi/10.1021/acs.energyfuels.8b01227.
DEMŠAR, Janez. Statistical comparisons of classifiers over multiple data sets. **Journal of Machine Learning Research**, [*s. l.*], v. 7, p. 1–30, 2006.

EBBS-PICKEN, Takiah; DA SILVA, Carlos M.; AMON, Cristina H. Design optimization methodologies applied to battery thermal management systems: A review. **Journal of Energy Storage**, [*s. l.*], v. 67, p. 107460, 2023. Disponível em: https://linkinghub.elsevier.com/retrieve/pii/S2352152X23008575.

FAN, Dongyan *et al.* Review of Machine Learning Methods for Steady State Capacity and Transient Production Forecasting in Oil and Gas Reservoir. **Energies**, [*s. l.*], v. 18, n. 4, p. 842, 2025. Disponível em: https://www.mdpi.com/1996-1073/18/4/842.

FAN, Jerome; UPADHYE, Suneel; WORSTER, Andrew. Understanding receiver operating characteristic (ROC) curves. **Canadian Journal of Emergency Medicine**, [*s. l.*], v. 8, n. 1, p. 19–20, 2006. Disponível em: https://www.cambridge.org/core/product/identifier/S1481803500013336/type/journal_article.

FEITOSA, Filipe Xavier. **Síntese de derivados do cardanol para desemulsificação de petróleo e inibição da precipitação asfaltênica**. 2018. 140 f. - Universidade Federal do Ceará (UFC), [*s. l.*], 2018. Disponível em: https://repositorio.ufc.br/bitstream/riufc/37067/7/2018_tese_fxfeitosa.pdf.

FERREIRA FILHO, Virgílio José Martins *et al.* Estimating water-in-oil emulsion viscosity of brazilian crude oils using machine learning techniques. **Rio Oil and Gas Expo and Conference**, [*s. l.*], v. 20, n. 2020, p. 448–449, 2020. Disponível em: https://biblioteca.ibp.org.br/riooilegas/pt-BR/search/40199?exp=.

GÉRON, Aurélien. **Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow**. [*S. l.: s. n.*], 2022.

ILIA ANISA, A. N.; NOUR, Abdurahman H. Affect of viscosity and droplet diameter on water-in-oil (w/o) emulsions: An experimental study. **World Academy of Science, Engineering and Technology**, [*s. l.*], v. 62, n. November, p. 691–694, 2010.

JURADO, Encarnación *et al.* Estimation of the distribution of droplet size, interfacial area and volume in emulsions. **Colloids and Surfaces A: Physicochemical and Engineering Aspects**, [*s. l.*], v. 295, n. 1–3, p. 91–98, 2007. Disponível em: https://linkinghub.elsevier.com/retrieve/pii/S092777570600656X.

KHATAEE, A. R.; KASIRI, M. B. Artificial neural networks modeling of contaminated water treatment processes by homogeneous and heterogeneous nanocatalysis. **Journal of Molecular Catalysis A: Chemical**, [*s. l.*], v. 331, n. 1–2, p. 86–100, 2010.

KHOUKHI, Amar. Hybrid soft computing systems for reservoir PVT properties prediction. **Computers and Geosciences**, [*s. l.*], v. 44, p. 109–119, 2012. Disponível em: https://linkinghub.elsevier.com/retrieve/pii/S0098300412001045.

KOKAL, Sunil. Crude-Oil Emulsions: A State-Of-The-Art Review. **SPE Production & Facilities**, [*s. l.*], v. 20, n. 01, p. 5–13, 2005. Disponível em: https://onepetro.org/PO/article/20/01/5/112495/Crude-Oil-Emulsions-A-State-Of-The-Art-Review.

KOKAL, Sunil. Crude Oil Emulsions: A State-of-the-Art Review. *In*: , 2002. **SPE Annual Technical Conference and Exhibition**. [*S. l.*]: SPE, 2002. Disponível em: https://onepetro.org/SPEATCE/proceedings/02ATCE/02ATCE/SPE-77497-MS/135984.

LAKE, W. Larry. **Petroleum Engineering Handbook, Volume 3**. [*S. l.: s. n.*], 2006.

LI, Changjun *et al.* An Experimental Study on the Viscosity of Water-in-Oil Emulsions. **Journal of Dispersion Science and Technology**, [*s. l.*], v. 37, n. 3, p. 305–316, 2016. Disponível em: http://www.tandfonline.com/doi/full/10.1080/01932691.2014.994218.

LIAN, Zhigang; GU, Xingsheng; JIAO, Bin. A novel particle swarm optimization algorithm for permutation flow-shop scheduling to minimize makespan. **Chaos, Solitons and Fractals**, [*s. l.*], v. 35, n. 5, p. 851–861, 2008. Disponível em: https://linkinghub.elsevier.com/retrieve/pii/S0960077906005388.

LIU, Wei *et al.* A systematic machine learning method for reservoir identification and production prediction. **Petroleum Science**, [*s. l.*], v. 20, n. 1, p. 295–308, 2023. Disponível em: https://doi.org/10.1016/j.petsci.2022.09.002.

LIU, Yigang *et al.* Experimental Study on Water-in-Heavy-Oil Droplets Stability and Viscosity Variations in the Dilution Process of Water-in-Heavy-Oil Emulsions by Light Crude Oil. **Energies**, [*s. l.*], v. 17, n. 2, p. 332, 2024. Disponível em: https://www.mdpi.com/1996-1073/17/2/332.

LV, Yuling *et al.* Oil-Water Two-Phase Flow with Three Different Crude Oils: Flow Structure, Droplet Size and Viscosity. **Energies**, [*s. l.*], v. 17, n. 7, p. 1573, 2024. Disponível em: https://www.mdpi.com/1996-1073/17/7/1573.

MA, Junwei *et al.* A comprehensive comparison among metaheuristics (MHs) for geohazard modeling using machine learning: Insights from a case study of landslide displacement prediction. **Engineering Applications of Artificial Intelligence**, [*s. l.*], v. 114, p. 105150, 2022. Disponível em: https://linkinghub.elsevier.com/retrieve/pii/S0952197622002664.

MAJNIK, Matjaž; BOSNIĆ, Zoran. ROC analysis of classifiers in machine learning: A survey. **Intelligent Data Analysis**, [*s. l.*], v. 17, n. 3, p. 531–558, 2013. Disponível em: https://journals.sagepub.com/doi/full/10.3233/IDA-130592.

MCKAY, M. D.; BECKMAN, R. J.; CONOVER, W. J. A Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output from a Computer Code. **Technometrics**, [*s. l.*], v. 21, n. 2, p. 239, 1979. Disponível em: https://www.jstor.org/stable/1268522?origin=crossref.

MENG, Jin *et al.* Hybrid data-driven framework for shale gas production performance analysis via game theory, machine learning, and optimization approaches. **Petroleum Science**, [*s. l.*], v. 20, n. 1, p. 277–294, 2023. Disponível em: https://linkinghub.elsevier.com/retrieve/pii/S1995822622002187.

MOHRI, Mehryarr; ROSTAMIZADEH, Afshin; TALWALKAR, Ameet. **Foundations of Machine Learning**. 2. ed. [*S. l.: s. n.*], 2018. Disponível em: https://mitpress.mit.edu/9780262039406/foundations-of-machine-learning/.

NANDI, B. K. *et al.* Treatment of oily wastewater using low cost ceramic membrane: Comparative assessment of pore blocking and artificial neural network models. **Chemical Engineering Research and Design**, [*s. l.*], v. 88, n. 7, p. 881–892, 2010. Disponível em: https://linkinghub.elsevier.com/retrieve/pii/S0263876209003190.

NAVID, Ali; KHALILARYA, Shahram; ABBASI, Mohammad. Diesel engine optimization with multi-objective performance characteristics by non-evolutionary Nelder-Mead algorithm: Sobol sequence and Latin hypercube sampling methods comparison in DoE process. **Fuel**, [*s. l.*], v. 228, p. 349–367, 2018. Disponível em: https://linkinghub.elsevier.com/retrieve/pii/S0016236118307762.

PANDEY, Yogendra Narayan *et al.* **Machine Learning in the Oil and Gas Industry: Including Geosciences, Reservoir Engineering, and Production Engineering with Python**. Berkeley, CA: Apress, 2020. Disponível em: https://link.springer.com/10.1007/978-1-4842-6094-4.

PEREIRA, Andréa da Silva. **Diagnóstico computacional e otimização operacional da unidade de dessulfurização industrial semi-seco (FGD-SDA) em plantas de geração termelétricas**. 2019. [*s. l.*], 2019.

ROMERO YANES, José Francisco *et al.* Addition of Non-endogenous Paraffins in Brazilian Crude Oils and Their Effects on Emulsion Stability and Interfacial Properties. **Energy & Fuels**, [*s. l.*], v. 33, n. 5, p. 3673–3680, 2019. Disponível em: https://pubs.acs.org/doi/10.1021/acs.energyfuels.8b02991.

RØNNINGSEN, Hans Petter. Correlations for predicting Viscosity of W/O-Emulsions based on North Sea Crude Oils. *In*: , 1995. **SPE International Symposium on Oilfield Chemistry**. [*S. l.*]: SPE, 1995. Disponível em: https://onepetro.org/SPEOCC/proceedings/95OCS/95OCS/SPE-28968-MS/57160.

ROSENDO, Matheus. **Um algoritmo de otimização por nuvem de partículas para resolução de problemas combinatórios**. 2010. 93 f. - Universidade Federal do Paraná (UFPR), [*s. l.*], 2010. Disponível em: https://acervodigital.ufpr.br/xmlui/bitstream/handle/1884/24862/dissertacao_matheus_rosendo.pdf?sequence=1&isAllowed=y.

SAAD, M. A. *et al.* An Overview of Recent Advances in State-of-the-Art Techniques in the Demulsification of Crude Oil Emulsions. **Processes**, [*s. l.*], v. 7, n. 7, p. 470, 2019. Disponível em: https://www.mdpi.com/2227-9717/7/7/470.

SALAGER, Jean-Louis; ISABEL BRICENO, María; LUIS BRACHO, Carlos. Heavy Hydrocarbon Emulsions Making Use of the State of the Art in Formulation Engineering. **Encyclopedic Handbook of Emulsion Technology**, [*s. l.*], p. 455–495, 2001.

SANTOS, Samara O.S. *et al.* Green Machine Learning: Analysing the Energy Efficiency of Machine Learning Models. *In*: , 2024. **Proceedings of the 35th Irish Systems and Signals Conference, ISSC 2024**. [*S. l.*]: IEEE, 2024. p. 1–6. Disponível em: https://ieeexplore.ieee.org/document/10603302/.

SANTOS, Hyago Braga dos *et al.* The use of machine learning models to estimate the viscosity of Brazilian water-in-crude oil emulsions. **Fuel**, [*s. l.*], v. 395, p. 135270, 2025. Disponível em: https://linkinghub.elsevier.com/retrieve/pii/S0016236125009950.

SCHRAMM. Colloid Rheology. *In*: SCHRAMM, Laurier L. (org.). **Emulsions, Foams, Suspensions, and Aerosols**. [*S. l.*]: Wiley, 2014. p. 209–258. Disponível em: https://onlinelibrary.wiley.com/doi/book/10.1002/9783527679478.

SHAKOURI, Sina; MOHAMMADZADEH-SHIRAZI, Maysam. Machine learning approaches for assessing stability in acid-crude oil emulsions: Application to mitigate formation damage. **Petroleum Science**, [*s. l.*], v. 22, n. 2, p. 894–908, 2025. Disponível em: https://doi.org/10.1016/j.petsci.2024.09.013.

SHEIKHOLESLAMI, Razi; RAZAVI, Saman. Progressive Latin Hypercube Sampling: An efficient approach for robust sampling-based analysis of environmental models. **Environmental Modelling and Software**, [*s. l.*], v. 93, p. 109–126, 2017. Disponível em: https://linkinghub.elsevier.com/retrieve/pii/S1364815216305096.

SILVA, Thiago Geraldo *et al.* AI Based Water-in-Oil Emulsions Rheology Model for Value Creation in Deepwater Fields Production Management. *In*: , 2021. **Proceedings of the Annual Offshore Technology Conference**. [*S. l.*]: OTC, 2021. Disponível em: https://onepetro.org/OTCONF/proceedings/21OTC/1-21OTC/D011S003R001/466766.

SJÖBLOM, Johan. **Encyclopedic handbook of emulsion technology**. [*S. l.*]: Marcel Dekker, 2001.

SJÖBLOM, Johan *et al.* Our current understanding of water-in-crude oil emulsions. Recent characterization techniques and high pressure performance. **Advances in Colloid and Interface Science**, [*s. l.*], v. 100–102, n. SUPPL., p. 399–473, 2003. Disponível em: https://linkinghub.elsevier.com/retrieve/pii/S0001868602000660.

SOUSA, M. *et al.* Performance of a biosurfactant produced by Bacillus subtilis LAMI005 on the formation of oil / biosurfactant / water emulsion: Study of the phase behaviour of emulsified systems. **Brazilian Journal of Chemical Engineering**, [*s. l.*], v. 31, n. 3, p. 613–623, 2014. Disponível em: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0104-66322014000300004&lng=en&tlng=en.

SOUSA, Ana M.; MATOS, Henrique A.; PEREIRA, Maria J. Properties of Crude Oil-in-Water and Water-in-Crude Oil Emulsions: A Critical Review. **Industrial and Engineering Chemistry Research**, [*s. l.*], v. 61, n. 1, p. 1–20, 2022. Disponível em: https://pubs.acs.org/doi/10.1021/acs.iecr.1c02744.

SPEIGHT, James G. **Petroleum Chemistry And Refining**. [*S. l.*]: CRC Press, 1997. Disponível em: https://www.taylorfrancis.com/books/9781482229349.

SPEIGHT, James G. **The Chemistry and Technology of Petroleum**. 5. ed. [*S. l.*]: CRC Press, 2014. Disponível em: https://www.taylorfrancis.com/books/9781439873908.

TALEBI, Roya *et al.* Application of soft computing approaches for modeling saturation pressure of reservoir oils. **Journal of Natural Gas Science and Engineering**, [*s. l.*], v. 20, p. 8–15, 2014. Disponível em: https://linkinghub.elsevier.com/retrieve/pii/S1875510014001462.
TCHOUKOV, Plamen *et al.* Role of asphaltenes in stabilizing thin liquid emulsion films. **Langmuir**, [*s. l.*], v. 30, n. 11, p. 3024–3033, 2014. Disponível em: https://pubs.acs.org/doi/10.1021/la404825g.

TERPILOWSKI, Maksim. scikit-posthocs: Pairwise multiple comparison tests in Python. **Journal of Open Source Software**, [*s. l.*], v. 4, n. 36, p. 1169, 2019. Disponível em: http://joss.theoj.org/papers/10.21105/joss.01169.

THEODORIDIS, Sergios. **Machine Learning: A Bayesian and Optimization Perspective, Second Edition**. [*S. l.*]: Elsevier, 2020. Disponível em: https://linkinghub.elsevier.com/retrieve/pii/C20190037727.

UMAR, Abubakar Abubakar *et al.* A review of petroleum emulsions and recent progress on water-in-crude oil emulsions stabilized by natural surfactants and solids. **Journal of Petroleum Science and Engineering**, [*s. l.*], v. 165, p. 673–690, 2018. Disponível em: https://linkinghub.elsevier.com/retrieve/pii/S0920410518301980.

VANDERPLAATS, Garret N. **Numerical Optimization Techniques for Engineering Design**. [*S. l.*]: Vanderplaats Research & Development Inc., 1999.

VIRTANEN, Pauli *et al.* SciPy 1.0: fundamental algorithms for scientific computing in Python. **Nature Methods**, [*s. l.*], v. 17, n. 3, p. 261–272, 2020. Disponível em: https://www.nature.com/articles/s41592-019-0686-2.

WANG, Weilun; CHAKRABORTY, Goutam; CHAKRABORTY, Basabi. Predicting the risk of chronic kidney disease (Ckd) using machine learning algorithm. **Applied Sciences (Switzerland)**, [*s. l.*], v. 11, n. 1, p. 1–17, 2021. Disponível em: https://www.mdpi.com/2076-3417/11/1/202.

WANG, Dongshu; TAN, Dapei; LIU, Lei. Particle swarm optimization algorithm: an overview. **Soft Computing**, [*s. l.*], v. 22, n. 2, p. 387–408, 2018. Disponível em: http://link.springer.com/10.1007/s00500-016-2474-6.

YONGUEP, Edith *et al.* Formation, stabilization and chemical demulsification of crude oil-in-water emulsions: A review. **Petroleum Research**, [*s. l.*], v. 7, n. 4, p. 459–472, 2022. Disponível em: https://linkinghub.elsevier.com/retrieve/pii/S2096249522000072.

ZHOU, Zhi Hua. **Machine Learning**. [*S. l.: s. n.*], 2021.

ZOLFAGHARI, Reza *et al.* Demulsification techniques of water-in-oil and oil-in-water emulsions in petroleum industry. **Separation and Purification Technology**, [*s. l.*], v. 170, p. 377–407, 2016. Disponível em: https://linkinghub.elsevier.com/retrieve/pii/S1383586616307195.

ANEXOS

Anexo A.1 – Planejamento experimental gerado pelo método LHS

Tabela A.1: Planejamento experimental realizado.

Ensaio	Temperatura LHS (°C)	Temperatura experimental (°C)	Concentração de NaCl (g/L)	Corte de Água (%)
1	40	40	240	30
2	41	40	172	23
3	43	45	123	47
4	44	45	49	7
5	45	45	181	21
6	46	45	160	43
7	48	50	115	45
8	50	50	189	25
9	51	50	111	39
10	53	55	168	29
11	54	55	213	27
12	56	55	144	11
13	57	55	107	37
14	59	60	66	3
15	60	60	222	35
16	62	60	103	5
17	64	65	41	31
18	68	70	226	9
19	70	70	176	1
20	71	70	94	13
21	75	75	131	41
22	76	75	82	33
23	77	75	45	15
24	79	80	164	17
25	80	80	74	19

Fonte: Próprio autor.

Anexo A.2 – Limites de busca, valores otimizados e descrição dos hiperparâmetros para os modelos de classificação e regressão

Tabela A.2.1: Hiperparâmetros para os modelos de classificação

Modelo	Hiperparâmetro	Limites de busca	Otimizado	Descrição
Árvore de decisão	max_depth	5 – 100	6	Profundidade máxima da árvore, limita o número de níveis para evitar overfitting.
	min_sample_split	2 – 10	2	Número mínimo de amostras necessárias para dividir um nó.
	criterion	gini, entropy	gini	Função para medir a qualidade da divisão.
Gradient Boosting	n_estimators	3 – 100	7	Número de árvores de boosting a serem construídas.
	learning_rate	0,01 – 0.9	0,8965	Taxa de aprendizado que controla a contribuição de cada árvore.
	max_depth	5 – 100	7	Profundidade máxima de cada árvore, controla a complexidade do modelo.
Floresta aleatória	n_estimators	3 – 100	32	Número de árvores na floresta.
	max_depth	5 – 100	8	Profundidade máxima de cada árvore na floresta.
	min_sample_split	2 – 10	3	Número mínimo de amostras para dividir um nó em uma árvore.
	max_samples	0,01 – 1	0,96	Fração ou número de amostras a serem usadas para treinar cada árvore.
Multi Layer Perceptron	activation	logistic, relu	logistic	Função de ativação das camadas ocultas
	solver	adam, sgd, lbfgs	lbfgs	Algoritmo de otimização
	alpha	0,0001 – 0,9	0,008662865732175392	Parâmetro de regularização L2 para penalizar pesos grandes.
	num_hidden_layers	1 – 4	3	Número de camadas ocultas na rede neural.
	num_units_per_layer	1 – 100	(5, 45, 33)	Número de neurônios em cada camada oculta.
	max_iter	-	7000	Número máximo de iterações para o treinamento da rede.

Fonte: Próprio autor.

Tabela A.2.2: Hipeparâmetros para os modelos de viscosidade.

Modelo	Hiperparâmetro	Limites de busca	Otimizado	Descrição
Regressão polinomial	polynomialfeatures_degree	2 – 11	2	Grau do polinômio para transformar as features, controla a complexidade do modelo.
XGBoost	n_estimators	3 – 100	9	Número de árvores de boosting a serem construídas.
	learing_rate	0,01 – 0,9	0,7782	Taxa de aprendizado que controla a contribuição de cada árvore.
	max_depth	5 – 100	9	Profundidade máxima de cada árvore, controla a complexidade do modelo.
	subsample	0,1 – 1	0,691	Fração dos dados usados para treinar cada árvore.
Floresta aleatória	n_estimators	3 – 100	77	Número de árvores na floresta.
	max_depth	5 – 100	25	Profundidade máxima de cada árvore na floresta.
	max_samples	0,01 – 1	0,876	Fração ou número de amostras a serem usadas para treinar cada árvore.
Multi-Layer Perceptron	activation	relu, tanh	relu	Função de ativação das camadas ocultas.
	solver	adam, sgd, lbfgs	lbfgs	Algoritmo de otimização
	num_hidden_layers	1 – 4	3	Número de camadas ocultas na rede neural.
	num_units_per_layer	5 – 100	(47, 13, 16)	Número de neurônios em cada camada oculta.
	alpha	0,0001 – 0,9	0,059482016720672375	Parâmetro de regularização L2 para penalizar pesos grandes.
	max_iter	-	7000	Número máximo de iterações para o treinamento da rede.

Fonte: Próprio autor.

Tabela A.2.3: Hipeparâmetros para os modelos de tamanho médio de gota

Modelo	Hiperparâmetro	Limites de busca	Otimizado	Descrição
Regressão polinomial	polynomialfeatures_degree	2 – 11	2	Grau do polinômio para transformar as features, controla a complexidade do modelo.
XGBoost	n_estimators	3 – 100	22	Número de árvores de boosting a serem construídas.
	learing_rate	0,01 – 0,9	0,9	Taxa de aprendizado que controla a contribuição de cada árvore.
	max_depth	5 – 100	7	Profundidade máxima de cada árvore, controla a complexidade do modelo.
	lambda	6 – 16	14,822	Parâmetro de regularização L2 para penalizar pesos das árvores.
	alpha	0,1 – 0,9	0,93	Parâmetro de regularização L1 para penalizar pesos das árvores.
Floresta aleatória	n_estimators	3 – 100	33	Número de árvores na floresta.
	max_depth	5 – 100	21	Profundidade máxima de cada árvore na floresta.
	max_samples	0,01 – 1	0,615	Fração ou número de amostras a serem usadas para treinar cada árvore.
Multi Layer Perceptron	activation	relu, tanh	relu	Função de ativação das camadas ocultas.
	solver	adam, sgd, lbfgs	lbfgs	Algoritmo de otimização
	num_hidden_layers	1 – 4	3	Número de camadas ocultas na rede neural.
	num_units_per_layer	1 – 100	(28, 71, 74)	Número de neurônios em cada camada oculta.
	alpha	0,0001 – 0,9	0,6719236890479191	Parâmetro de regularização L2 para penalizar pesos grandes.
	max_iter	-	7000	Número máximo de iterações para o treinamento da rede.

Fonte: Próprio autor.