UNIVERSIDADE FEDERAL DE ITAJUBÁ

PROGRAMA DE PÓS GRADUAÇÃO EM
ENGENHARIA DE PRODUÇÃO

Impacto das variáveis meteo-oceanográficas e do teor de óleos e graxas na formação de feições oleosas durante o processamento primário de petróleo

Estevão Luiz Romão

Itajubá, junho de 2022

UNIVERSIDADE FEDERAL DE ITAJUBÁ

PROGRAMA DE PÓS GRADUAÇÃO EM
ENGENHARIA DE PRODUÇÃO

Estevão Luiz Romão

Impacto das variáveis meteo-oceanográficas e do teor de óleos e graxas na formação de feições oleosas durante o processamento primário de petróleo

Tese submetida ao Programa de Pós-Graduação em Engenharia de Produção da Universidade Federal de Itajubá como parte dos requisitos para a obtenção do título de Doutor em Ciências em Engenharia de Produção.

Área de concentração: Engenharia de Produção

Orientador: Prof. Pedro Paulo Balestrassi, Dr.

Coorientador: Aloisio Euclides Orlando Jr., Dr.

Itajubá, junho de 2022

UNIVERSIDADE FEDERAL DE ITAJUBÁ

PROGRAMA DE PÓS GRADUAÇÃO EM
ENGENHARIA DE PRODUÇÃO

Estevão Luiz Romão

Impacto das variáveis meteo-oceanográficas e do teor de óleos e graxas na formação de feições oleosas durante o processamento primário de petróleo

Tese aprovada por banca examinadora no dia 10 de junho de 2022, conferindo ao autor o título de ***Doutor em Ciências em Engenharia de Produção***.

Banca examinadora:

Prof. Dr. Claudimar Pereira da Veiga (UFPR)

Prof. Dr. Wesley Vieira da Silva (UFAL)

Prof. Dr. Anderson Paulo de Paiva (UNIFEI)

Prof. Dr. Antonio Fernando Branco Costa (UNIFEI)

Dr. Aloisio Euclides Orlando Jr. (Coorientador - PETROBRAS)

Prof. Dr. Pedro Paulo Balestrassi (Orientador)

Itajubá, junho de 2022

DEDICATÓRIA

A minha querida noiva, Daniela, pelo companheirismo de sempre.

AGRADECIMENTOS

A Deus por mais essa oportunidade!

Ao meu grande amigo e orientador Pedro Paulo Balestrassi pelo enorme apoio e contribuição para o desenvolvimento desse trabalho. Ao meu amigo e coorientador Aloisio Euclides Orlando Jr. por tornar possível o desenvolvimento desse trabalho. Ao meu amigo professor Anderson Paulo de Paiva por todas as suas valiosas contribuições durante todo o desenvolvimento dessa pesquisa.

A minha mãe, Orceli e meu pai Juarez que sempre me apoiaram na minha carreira acadêmica e também ao meu padrasto Warlen por sempre me incentivar.

Agradeço imensamente à minha querida noiva Daniela por todo apoio, carinho e companheirismo, estando sempre ao meu lado me dando muita força e não me deixando nunca desanimar.

À minha grande amiga Simone que sempre me ajudou em todos os momentos desse doutorado e a quem serei eternamente grato pela convivência e trabalho em equipe que desenvolvemos juntos, inclusive durante a elaboração e conclusão desse trabalho.

A todos os meus familiares que sempre estiveram presentes em minha vida, sem eles nunca teria chegado aonde cheguei, em especial a minha tia Mirtes e minha prima Lívia. Agradeço também ao meu sogro José Mauro e minha sogra Vandelice pelo carinho e apoio.

Muito obrigado aos meus grandes amigos da UNIFEI: Andreza, Alexandre, Aline, Taynara, Vinícius, Gustavo, Franco e Natália. Em especial ao Fabrício e ao Eduardo que tanto me auxiliaram durante esse doutorado.

Aos meus queridos amigos da casa espírita Fé, Esperança e Caridade, a todos o meu muito obrigado!

Agradeço imensamente aos professores que tanto contribuíram para a minha formação: Pedro Paulo Balestrassi, Anderson Paiva, e aos membros da banca examinadora, pelos seus comentários que muito auxiliaram na melhoria deste trabalho.

Agradeço imensamente ao apoio financeiro e estrutural fornecido pela CAPES, CNPq, PETROBRAS e UNIFEI que tornou possível a elaboração desse trabalho.

EPÍGRAFE

Ninguém é tão ignorante que não tenha algo a ensinar. Ninguém é tão sábio que não tenha algo a aprender. Blaise Pascal

RESUMO

O surgimento de feições oleosas no oceano é um desafio para as empresas que realizam o processamento primário de óleo em plataformas *offshore*. Após a separação do gás, óleo e água presentes no petróleo bruto, parte da água é devolvida aos oceanos com um certo teor de óleos e graxas (TOG). O valor do TOG e de variáveis meteo-oceanográficas, tais como: direção do vento (DV), intensidade do vento (IV), direção da corrente (DC), intensidade da corrente (IC), direção da onda (DO) e o período de pico primário (PP), criam cenários que podem favorecer ou dificultar o aparecimento de feições oleosas. No Brasil, essas feições podem levar a sanções para as empresas caso ultrapassem 500 metros de extensão. Diante disso, o presente trabalho realiza um estudo sobre como tais variáveis influenciam a probabilidade de ocorrência e detecção de feições oleosas via satélite, bem como a sua extensão. Utilizaram-se técnicas de *machine learning* (*random forest*, *k-nearest neighbors*, redes neurais artificiais, regressão logística e *support vector machines*), análise fatorial, *design of experiments* (DOE) e o algoritmo de otimização *desirability*. As principais conclusões do estudo foram: (i) o *random forest* superou os demais classificadores analisados e um modelo com área sob a curva de Característica de Operação do Receptor (curva ROC) de 0,93 foi obtido; (ii) a metodologia utilizada, combinando os classificadores com as técnicas anteriormente mencionadas mostrou-se satisfatória; (iii) quanto maior os valores de IV, DV e IC, menor a probabilidade de ocorrência e detecção de manchas de óleo, sendo que quanto maiores os valores de TOG, PP, DO e DC maior esta probabilidade; (iv) variáveis como IC e TOG contribuem positivamente para aumentar a extensão das manchas de óleo, enquanto altos valores de DV, IV e PP reduzem a extensão das feições.

Palavras-chave: Processamento primário de petróleo; variáveis meteo-oceanográficas; Teor de óleos e graxas; Técnicas de aprendizado de máquina; Planejamento de experimentos.

ABSTRACT

The appearance of oil sheens in the ocean is a challenge for companies that perform primary oil processing on offshore platforms. After the separation of the gas, oil and water that are present in crude oil, part of the water is returned to the oceans with a certain content of oils and greases. The value of the total oil and greases (TOG) associated with the values of metoceanographic variables such as: wind direction (WD), wind speed (WS), current direction (CD), current speed (CS), wind wave direction (WWD) and peak period (PP) create scenarios that favor or hinder the appearance of oil sheens. In Brazil, these oil sheens can lead to sanctions for companies if they exceed 500 meters in length. In view of this, the present work conducts a study about how such variables influence the probability of occurrence and detection of oil sheens via satellite, as well as their extent, applying machine learning techniques (*random forest*, *k-nearest neighbors*, artificial neural networks, logistic regression, and *support vector machines*), factor analysis, *design of experiments* (DOE) and the optimization algorithm *desirability*. The main conclusions of the study were: (i) random forest outperformed the other analyzed classifiers and a model whose area under the Receiver Operating Characteristic Curve (ROC curve) was 0.93 was achieved; (ii) the methodology used, combining the classifiers with the aforementioned techniques proved to be satisfactory; (iii) the higher the values of WS, WD and CS, the lower the probability of occurrence and detection of oil sheens, whereas the higher the values of TOG, PP, WWD and CD the higher the value of this probability; (iv) variables such as CS and TOG contribute positively to increasing the extension of the oil sheens, while high values of WD, WS and PP reduce the extension of the features.

Keywords: Primary oil processing; Meteoceanographic variables; Total oil and grease; Machine learning techniques; Design of experiments.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE QUADROS

SUMÁRIO

1. INTRODUÇÃO

A presente seção tem como objetivo apresentar o contexto no qual a presente pesquisa foi desenvolvida, os objetivos gerais e específicos que procura atingir, a justificativa e relevância do trabalho, as contribuições e delimitações da pesquisa, bem como a estrutura que o trabalho apresenta.

1.1. Contexto da pesquisa

O petróleo possui significativa participação na vida dos homens desde tempos remotos. Era utilizado pelos Egípcios para pavimentar estradas, embalsamar os mortos e mesmo nas construções de pirâmides. Os gregos, fenícios, romanos e os babilônicos também fizeram uso do petróleo com diferentes finalidades (TRIGGIA *et al.*, 2001). Atualmente, o petróleo possui uma imensa importância na matriz energética mundial, podendo ser considerado uma das fontes energéticas de maior importância para a economia dos países (KLEMZ *et al.*, 2021).

De acordo com Hsu e Robinson (2006) o petróleo é considerado um combustível fóssil devido ao fato de ser originado a partir dos corpos de organismos que habitaram o planeta há muitos anos. Ainda segundo o autor, ao morrerem, diversas criaturas têm seus restos acumulados no fundo de lagos e mares juntamente com outros sedimentos. Tais depósitos foram transformados em rochas sedimentares devido à pressão, ao calor e à ação bacteriana. Em seguida, a matéria orgânica foi transformada em simples compostos químicos, tais como hidrocarbonetos, dióxido de carbono, água, entre outros.

Dessa forma, o petróleo é constituído por centenas de compostos de hidrocarbonetos, ou seja, compostos que são formados por átomos de carbono e hidrogênio. Eles podem variar desde os mais simples, com apenas um átomo de carbono, até os mais complexos com cerca de 300 átomos de carbono, como parafinas ou isômeros de parafinas (JONES; PUJADÓ, 2006).

Em se tratando de sua extração, o petróleo pode ser obtido por meio de processos de recuperação primário, secundário ou terciário (JAVADI *et al.*, 2021). No primeiro estágio o petróleo é obtido por meio da própria energia do reservatório, que é função da pressão a qual os fluidos estão submetidos. Alguns mecanismos naturais mantém elevada a pressão interna desses reservatórios, dentre eles podem-se citar os influxos do gás em

solução, da capa de gás e da água, a expansão do fluido e da rocha, além da drenagem da gravidade (JAVADI *et al.*, 2021; TRIGGIA *et al.*, 2001).

Já no processamento de recuperação secundário, é preciso fornecer energia extra ao reservatório como forma de obtenção do petróleo. Geralmente essa energia é fornecida por meio de injeção de água ou gás (JAVADI *et al.*, 2021; LUO *et al.*, 2017). Enquanto que o processo de recuperação terciário, também conhecido como recuperação de petróleo avançada (*Enhanced Oil Recovery* – EOR), engloba a utilização de forças externas e substâncias de modo a interagirem térmica, física, biológica e quimicamente, com os reservatórios, possibilitando assim maiores taxas de recuperação (AL ADASANI; BAI, 2011).

Nesse contexto, a injeção de gás tem tido maiores aplicações, alguns incentivos a essa prática estão relacionados à falta de mercado para exportar gás produzido de campos localizados remotamente. Além disso, o gás é mais fácil de ser injetado quando comparado com métodos EOR de água ou químicos, uma vez que sua injeção não requer desenvolvimento da miscibilidade para ter sucesso, embora possa ajudar. Por outro lado, pode sofrer de baixa eficiência de varredura devido à heterogeneidade geológica, digitação viscosa, entre outros. Portanto, os processos de recuperação avançada à base de gás requerem maior controle para melhorar a eficiência da varredura (MOGENSEN; MASALMEH, 2020).

Após sua extração, o petróleo cru passa pelo processamento primário, que consiste em separar as fases, isto é, óleo, água e gás das impurezas em suspensão. Em seguida os hidrocarbonetos são condicionados para serem transferidos às refinarias e a água produzida é tratada para ser reinjetada ou descartada (TRIGGIA *et al.*, 2001). Plataformas de processamento primário de petróleo *offshore* (localizadas no mar), alvo do presente estudo, necessitam de grande quantidade de energia para extrair, processar e transportar o petróleo, que geralmente é suprida por meio de produção local de gás e turbinas a gás, utilizadas para geração de eletricidade demandada por bombas e compressores (CARRANZA SÁNCHEZ; DE OLIVEIRA, 2015).

A separação entre as fases do petróleo pode ser realizada em separadores bifásicos ou trifásicos, atuando em série ou em paralelo. Nos equipamentos bifásicos, ocorre a separação gás/líquido, enquanto que nos trifásicos ainda ocorre a separação do óleo e da água. Nesse contexto, os hidrociclones e a flotação são procedimentos comumente

utilizados pela indústria do petróleo para separar óleo e a água. A flotação busca recuperar o óleo por meio da ação gravitacional e os hidrociclones aceleram esse processo, no qual uma força centrípeta age de forma a forçar os componentes mais pesados (água e sólidos) contra as suas paredes (TRIGGIA *et al.*, 2001).

De acordo com Klemz *et al.*, (2021), estima-se que sejam produzidos dezenas de milhões de barris de água por dia ao redor do mundo, e os processos para tratamento dessa água tem se tornado um desafio para a indústria petrolífera. Assim, a água descartada que retorna ao oceano é acrescida de óleos e graxas, e quando o nível desses compostos é muito elevado pode ser extremamente prejudicial ao meio ambiente. Nesse sentido, o teor de óleos e graxas (TOG) é utilizado como indicador cujo limite deve ser monitorado.

Segundo a RESOLUÇÃO CONAMA n° 393[79], de 8 de agosto de 2007 ("CONAMA Resolution No. 393/2007", 2007), que dispõe sobre a prevenção, o controle e a fiscalização da poluição causada por lançamento de óleo e outras substâncias nocivas ou perigosas em águas sob jurisdição nacional, em seu Art. 5o, estabelece que o TOG mensal, obtido por meio da média aritmética dos valores diários, deverá ser inferior a 29 mg/L, e os valores diários não deverão exceder 42 mg/L. Enquanto isso, o Art. 15o estabelece que o não cumprimento do disposto nesta Resolução sujeitará os infratores às sanções previstas pela legislação vigente. É importante destacar que esses valores de TOG devem ser obtidos em medições terrestres por gravimetria.

Além disso, a mesma resolução estabelece que a água descartada não deve alterar características do mar para além da zona de mistura, limitada a 500 metros do ponto de descarte. Dessa forma, o aparecimento de feições oleosas com extensão superior a esse valor acarretará penalidades à empresa que realiza o processamento primário de petróleo. Nesse contexto, é importante destacar que a zona de mistura pode ser entendida como a região do corpo receptor onde ocorre a diluição inicial do efluente.

Aliado ao TOG da água produzida no processamento primário de petróleo, variáveis relacionadas às condições temporais do ambiente também podem exercer influência significativa sobre o aparecimento de feições oleosas com extensão superior a 500 metros. Essas variáveis são denominadas meteo-oceanográficas, sendo que as seguintes foram consideradas no presente estudo: direção do vento (°), intensidade do vento (m/s), direção da corrente (°), intensidade da corrente (m/s), direção da onda (°) e período de pico primário (s).

Alguns trabalhos disponíveis na literatura dispõem a respeito da influência de variáveis meteo-oceanográficas sobre a trajetória e espessura das machas de óleo em água. Sabe-se que variáveis meteo-oceanográficas relacionadas à corrente, vento e ondas podem ter uma influência significativa na trajetória do óleo derramado na água. Dentre essas variáveis, o vento e a corrente do mar apresentaram as maiores influências nos modelos de previsão da trajetória do óleo (PISANO *et al.*, 2016). O efeito do vento na dispersão natural do óleo na água do mar também foi avaliado, constatando que a velocidade do vento e as propriedades do óleo, como densidade e tensão superficial, têm grande influência na espessura da camada de óleo. (ZATSEPA *et al.*, 2018).

Além disso, o estudo desenvolvido por (DANESHGAR ASL *et al.*, 2017) demonstraram que altos valores de velocidade do vento, maiores que 7 m/s, contribuem para o desaparecimento de manchas de óleo. Por outro lado, maiores valores associados à intensidade de corrente foram responsáveis por gerar manchas de óleo com maior extensão e consequentemente menor espessura.

A partir disso, surge o problema, alvo da presente pesquisa, que é entender quais condições favorecem o aparecimento de uma feição oleosa, e avaliar a capacidade de prever a ocorrência de tais feições. É necessário estabelecer o nível para o qual o valor de TOG (mg/L) deve ser levado a fim de reduzirem a probabilidade de formação da feição oleosa, uma vez que é a única variável que pode ser controlada dentre aquelas consideradas no estudo.

1.2. Objetivos

1.2.1. Objetivo geral

Em suma, o objetivo principal do trabalho é verificar o impacto das variáveis meteo-oceanográficas e do TOG Espectrofotométrico na formação de feições oleosas que possuem extensão superior a 500 metros. Para tal, será desenvolvido um método para previsão da ocorrência e detecção da feição com base nas variáveis anteriormente mencionadas. Esse modelo deverá possibilitar a compreensão do efeito dessas variáveis no aparecimento da feição oleosa, bem como em sua extensão. Para tal, é necessário obter um valor de probabilidade associado a essa previsão binária (presença ou ausência de feição) que possa auxiliar os tomadores de decisão do processo.

Assim, é importante destacar que a presente tese visa a defender a hipótese de que a probabilidade do surgimento de feições oleosas não depende apenas do valor de TOG, mas também das as variáveis meteo-oceanográficas, possuindo, essas últimas, um papel de extrema importância nesse cenário. Espera-se conseguir identificar quais os níveis que as variáveis consideradas devem assumir a fim de reduzir o valor da probabilidade de ocorrência de feição.

1.2.2. Objetivos específicos

Em termos específicos, o presente trabalho objetiva:

- Avaliar se existem cenários meteo-oceanográficas que favorecem o aparecimento de feições oleosas, bem como avaliar o impacto relacionado à variável TOG Espectrofotométrico.
- Desenvolver modelos por meio de diversos métodos de classificação disponíveis em Python e compará-los a fim de descobrir o método mais indicado para a solução do presente problema.
- Avaliar se a utilização de *ensembles* como classificadores para a previsão de ocorrência de feições oleosas é mais indicada do que os classificadores individuais.
- Utilizar técnicas de estatística multivariada para auxiliar no processo de previsão, uma vez que as variáveis podem apresentar correlações significativas entre si.
- Modelar a probabilidade de ocorrência de feição em função das variáveis meteo-oceanográficas e do TOG Espectrofotométrico.
- Modelar a extensão da feição oleosa em função das variáveis meteo-oceanográficas e do TOG Espectrofotométrico.

1.3. Justificativa

A presente pesquisa se justifica pela lacuna existente na literatura em termos de um estudo de como as diversas variáveis meteo-oceanográficas, mencionadas anteriormente, impactam a formação e a extensão de feições oleosas durante o processamento primário de petróleo. A variável TOG certamente está envolvida no processo aparecimento e detecção de feições oleosas, uma vez que é ela que indica o teor de óleos e graxas na água descartada, entretanto não foram encontrados estudos que avaliam o TOG obtido por meio

espectrofotometria de absorção molecular (TOG Espectrofotométrico) no aparecimento das feições.

Além disso, existe a necessidade de elaboração de um modelo capaz de prever satisfatoriamente o risco de ocorrência dessas feições. Geralmente, nas plataformas *offshore* de processamento primário de petróleo, as variáveis meteo-oceanográficas e o TOG Espectrofotométrico são monitorados constantemente, assim existe a viabilidade de obtenção de um modelo de classificação capaz de fornecer o risco de ocorrência e detecção de uma feição oleosa.

Essa capacidade de reagir preventivamente à formação de feições oleosas é de extrema importância, pois podem ser extremamente prejudicial à vida marinha (MORANDIN; O'HARA, 2016; O'HARA; MORANDIN, 2010). Busca-se, por meio de pesquisas como essa atingir o chamado desenvolvimento sustentável que é definido no Relatório de Brundtland, em 1987, como sendo o tipo de desenvolvimento que atende às necessidades do presente sem comprometer a capacidade das gerações futuras de atender às suas próprias necessidades. Dessa forma, a possibilidade de evitar danos ao meio ambiente durante o processamento primário de petróleo é de extrema importância no atual cenário da humanidade.

Em termos acadêmicos, é possível observar a relevância de publicações relacionadas a petróleo após uma busca na base *Web of Science* em artigos com o termo *petroleum* no título conforme mostrado na Figura 1.

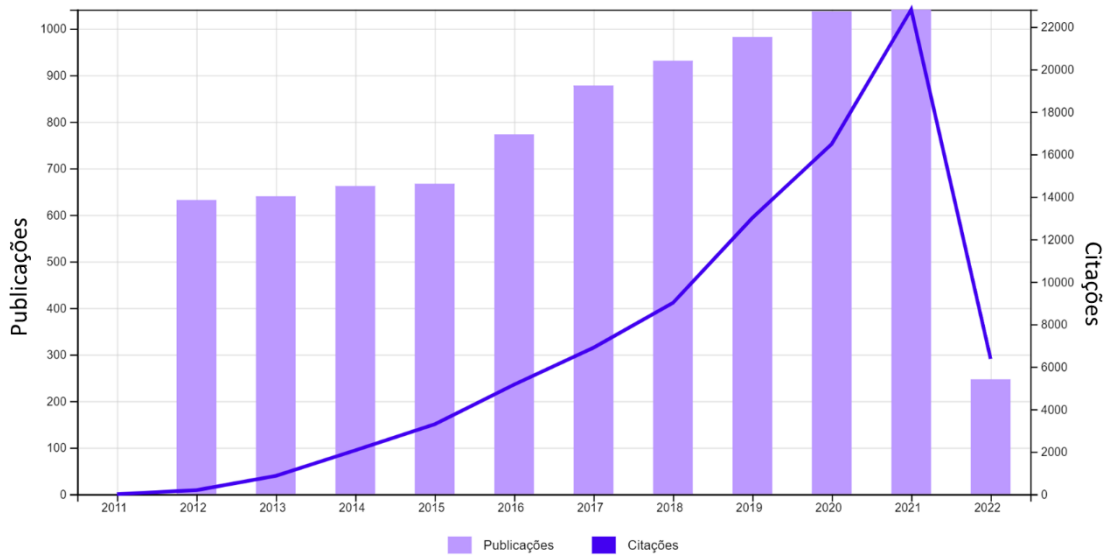


Figura 1. Publicações e citações de artigos com o termo '*petroleum*' no título na *Web of Science*

Foram obtidas um total de 8.490 publicações no período avaliado de 2012 até maio de 2022. Além disso, também é possível observar um crescimento significativo em relação ao número de citações mostrado pela linha azul com tendência de crescimento.

As principais áreas relacionadas a esse tema podem ser visualizadas na Figura 2. Pode-se observar uma grande predominância das áreas relacionadas à energia, ciências ambientais, engenharia química, engenharia de petróleo, entre outras. Entretanto, o termo '*total oil and grease*' (TOG) tem sido pouco explorado, uma vez que apresentou apenas 4 resultados nos últimos 10 anos na *Web of Science*, demonstrando uma lacuna a ser explorada, uma vez que pesquisas relacionadas a petróleo possuem tanto interesse na comunidade científica.

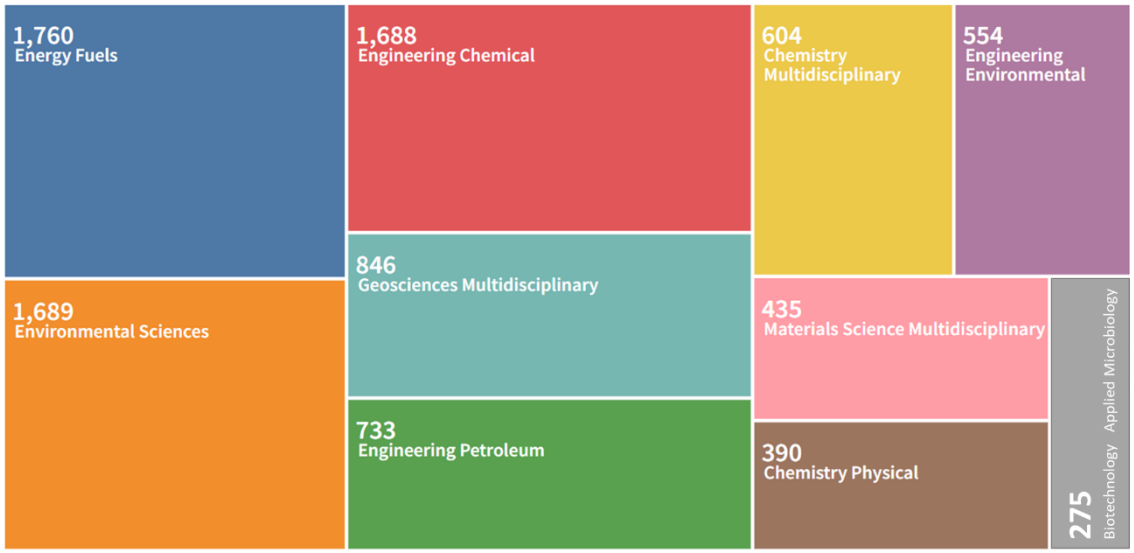


Figura 2. Quantidade de publicações relacionadas ao termo '*petroleum*' separados por área de conhecimento segundo a *Web of Science*

O termo *machine learning* também foi utilizado para uma busca na base *Web of Science* e nos últimos 10 anos foram encontradas mais de 65 mil publicações. Ao buscar mais especificamente pelo termo *random forest*, método utilizado tanto em problemas de classificação como regressão que foi amplamente utilizado na presente pesquisa, foram encontradas 5.988 publicações, o que demonstra a grande utilização do método. Esse resultado pode ser observado na Figura 3.

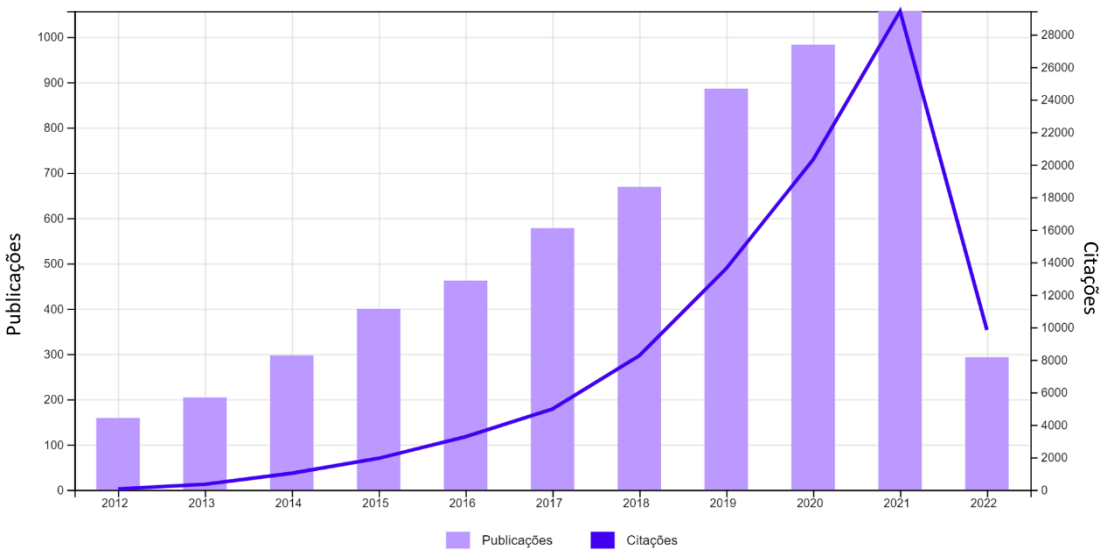


Figura 3. Publicações e citações de artigos com o termo '*random forest*' no título na *Web of Science*

As principais áreas relacionadas a utilização do método *random forest* podem ser observadas na Figura 4. Conclui-se que as principais áreas são relacionadas à engenharia elétrica, ciências da computação, sistemas de informação, entre outras.

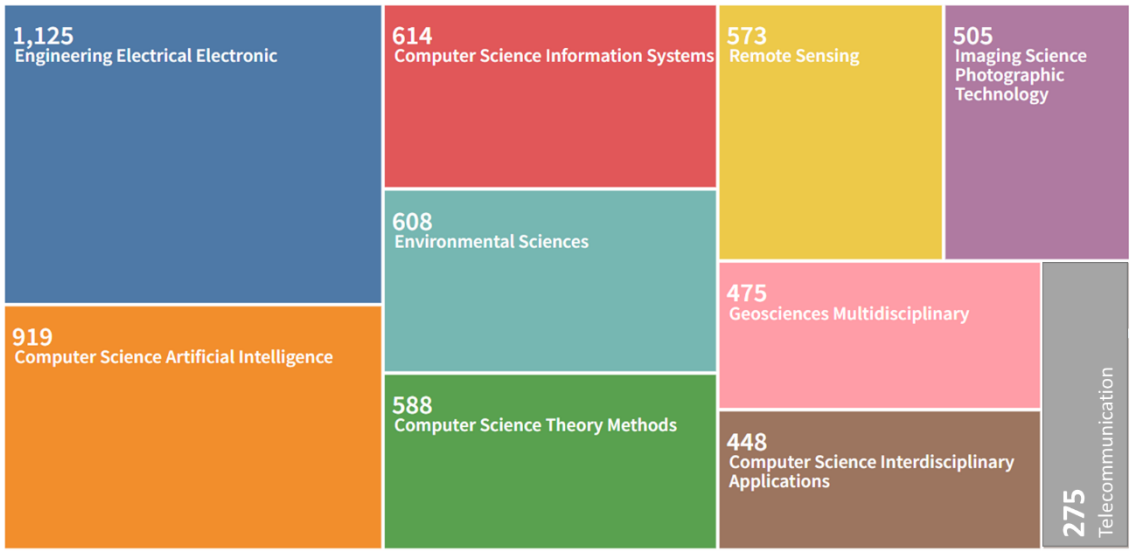


Figura 4. Quantidade de publicações relacionadas ao termo '*random forest*' separados por área de conhecimento segundo a *Web of Science*

Termos como *oil sheen* e *oil film* foram buscados na mesma base, entretanto poucos resultados são exibidos, demonstrando que esse tipo de estudo ainda precisa de ser explorado em pesquisas atuais.

1.4. Contribuições esperadas

Uma vez que o processamento primário de petróleo é uma atividade essencial para o funcionamento de diversos setores da sociedade atual, espera-se que com a presente pesquisa essa atividade possa ser otimizada de modo que cada vez menos impactos sejam causados ao meio ambiente.

A partir das conclusões deste trabalho, espera-se conhecer o impacto das variáveis meteo-oceanográficas e do TOG no processo de formação e detecção de feições oleosas. Na prática, espera-se que seja possível elaborar um modelo preditivo a ser utilizado em plataformas de processamento primário *offshore*. Um modelo para previsão da ocorrência e detecção de feições oleosas poderia ser aplicado de modo auxiliar a plataforma a reduzir valores de TOG em momentos nos quais as condições meteo-oceanográficas são favoráveis à geração de feição com mais de 500 metros, podendo causar danos à vida marinha na região.

Espera-se também, contribuir com uma metodologia que faz uso de algoritmos de classificação, métodos multivariados e algoritmos de otimização utilizados conjuntamente. Assim, a metodologia aplicada nessa pesquisa não deve ser limitada a problemas relacionados a processamento primário de petróleo, mas a todo problema de classificação em que os efeitos de possíveis variáveis preditoras sobre determinada resposta ainda não são conhecidos ou não foram amplamente explorados em estudos anteriores.

1.5. Delimitações da pesquisa

Em relação às delimitações do trabalho, convém ressaltar que se utilizaram apenas cinco classificadores durante o desenvolvimento da pesquisa: *random forest*, k-vizinhos mais próximos, redes neurais artificiais, regressão logística binária e máquina de vetores de suporte, além da combinação dos 3 que, dentre eles, apresentaram melhores valores das métricas de avaliação. A utilização de um classificador, ou de uma combinação de classificadores, que apresenta maior precisão poderia melhorar a interpretação dos efeitos das variáveis, bem como ser utilizada durante a previsão de detecção e ocorrência de feição.

A variável altura significativa, que pode ser definida como a média relacionada aos 33% maiores valores de altura de onda, não foi considerada nas análises por apresentar

muitos valores faltantes. A inclusão e análise dessa variável no modelo é importante, pois além de permitir a análise do efeito dessa variável, também poderá aumentar a precisão do modelo. Além disso variáveis relacionadas ao processamento primário de petróleo, como por exemplo, pressão no flotador, vazão de saída, temperatura da água descartada, utilização de produtos químicos, entre outras, também não foram englobadas nesse estudo.

O TOG Gravimétrico, utilizado a nível fiscal e mensurado em terra, e devido à metodologia empregada na sua medição, é inviável sua realização na plataforma. Isso implicaria na impossibilidade de agir preventivamente sobre a ocorrência da feição. Dessa forma, no presente estudo, utilizou-se outra medição que é realizada a bordo: o TOG Espectrofotométrico.

Por fim, os dados trabalhados foram obtidos de uma única uma plataforma de processamento primário de petróleo.

1.6. Estrutura do trabalho

O presente trabalho está estruturado de forma a facilitar a compreensão do leitor em relação aos temas abordados. Dessa forma, a seção 2 apresenta uma revisão de literatura na qual são abordados os tópicos essenciais sobre processamento primário de petróleo e efeito iridescente, discussões necessárias para possibilitar a compreensão do contexto em que esse trabalho está inserido.

Em seguida, ainda na seção 2, são apresentados os principais conceitos relacionados a problemas de classificação e aos principais métodos considerados no estudo: *Random Forest*, K-vizinhos mais próximos, redes neurais artificiais, regressão logística binária, máquina de vetores de suporte. Por fim, essa seção ainda aborda a metodologia de superfície de resposta, utilizada para elaborar um modelo para a probabilidade de ocorrência e detecção de feição e também disserta a respeito da análise fatorial, técnica utilizada para reduzir a dimensionalidade do problema e também representar as variáveis originais por meio de *scores* de fatores não correlacionados.

Os materiais e métodos utilizados para a elaboração dessa pesquisa são apresentados na seção 3. Essa seção está dividida em 3 tópicos, são eles: caracterização da pesquisa; coleta de dados, descrevendo a forma como os dados são obtidos ressaltando a

instrumentação em plataformas *offshore*; e uma descrição das metodologias utilizadas a análise dos dados.

A seguir, na seção 4, são apresentados os principais resultados e discussões divididos em tópicos a fim de facilitar a compreensão do leitor. Desenvolveram-se algumas análises preliminares, assim como o passo a passo da metodologia proposta na seção 3, sendo também apresentados os resultados relacionados à análise de risco, ao desempenho dos classificadores, ao modelo para a probabilidade de ocorrência e detecção de feições oleosas e ao modelo para a extensão da feição oleosa.

Finalmente a seção 5 apresenta as conclusões do trabalho, bem como as contribuições e sugestões para trabalhos futuros. A seção 6 apresenta as referências bibliográficas utilizadas para embasar o desenvolvimento dessa pesquisa. Os testes t pareado realizados no Minitab são apresentados no Apêndice deste trabalho. Os conjuntos de dados utilizados, tanto para ocorrência de feição como para extensão da feição são apresentados nos Anexos A. Por fim, os artigos publicados em periódicos durante o doutorado estão apresentados no Anexo B.

2. REVISÃO DA LITERATURA

Nesta seção serão apresentados os principais conceitos a serem compreendidos a fim de facilitar o entendimento das próximas seções. Nas duas primeiras subseções serão abordados os conceitos relacionados ao contexto da pesquisa, ou seja, processamento primário de petróleo e efeito iridescente. As demais subseções apresentam discussões a respeito dos algoritmos de classificação (*random forest*, k-vizinhos mais próximos, redes neurais artificiais, regressão logística binária e máquina de vetores de suporte), assim como a metodologia de superfície de resposta e análise fatorial, utilizadas ao longo da aplicação da metodologia proposta na pesquisa.

2.1. Processamento primário de petróleo

Uma plataforma de processamento primário de petróleo do tipo *offshore* possui como principal objetivo a separação dos componentes gás, óleo e água contidos no petróleo bruto extraído dos reservatórios marinhos. Segundo Carranza Sánchez e De Oliveira (2015), as plataformas *offshore* podem ser dividas em plantas de processamento e utilitárias. As últimas, segundo os autores, são dão suporte às operações de energia, segurança e serviços na plataforma, sendo que os processos de conversão de energia compreendem:

- geração de energia e compressão de ar;
- operações de segurança, incluindo sistemas de água de incêndio e óleo diesel para geradores de emergência e serviço de água do mar;
- instalações de reabastecimento de helicópteros, água potável e sistema de drenagem.

Por outro lado, as plataformas de processamento são responsáveis pelo processamento de hidrocarbonetos, exportação de petróleo e gás, assim como injeção de gás. Dessa forma, esse tipo de plataforma engloba diversos processos usados para separar o petróleo e refinar os produtos para uma qualidade adequada. O Quadro 1 mostra os principais processos realizados nesse tipo de plataforma de acordo com o tipo de processamento realizado pela plataforma *offshore*, ou seja, de óleo, de gás ou de água (CARRANZA SÁNCHEZ e DE OLIVEIRA, 2015).

Quadro 1. Atividades desenvolvidas nos diferentes tipos de processamentos realizados em plataformas *offshore*

Tipo de processamento	Processos
Processamento de óleo	- Redução da pressão na cabeça do poço; - Redistribuição do petróleo bruto no aquecedor de produção; - Separação nos separadores de produção; - Medição das quantidades exatas de gás água e condensado produzidos no separador de teste; - Condicionamento e bombeamento do petróleo bruto.
Processamento de gás	- Remoção de líquido antes do primeiro estágio de compressão de gás no purificador de entrada; - Compressão de gás para exportação; - Reinjeção ou operações de elevação de gás, queima, desidratação de gás e adoçamento de gás.
Processamento de água	- Tratamento da água produzida durante o processo de separação para ser lançada no mar ou injetada em um reservatório para manutenção da pressão.

Fonte: Adaptado de Carranza Sánchez e De Oliveira (2015)

A Figura 5 demonstra o fluxograma de processamento primário de fluidos de acordo com (TRIGGIA *et al.*, 2001), apresentando os principais componentes de uma unidade de processamento de petróleo.

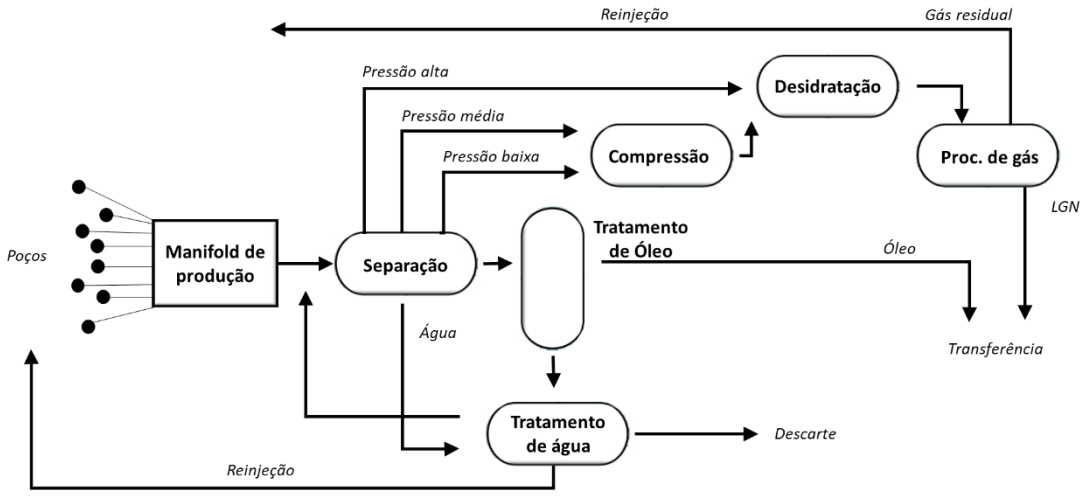


Figura 5. Processamento primário de fluidos (adaptado de (TRIGGIA *et al.*, 2001))

É importante destacar que o *manifold* de produção consiste em um conjunto de válvulas e outros acessórios capazes de direcionar a produção advinda de vários poços

para um duto coletor, sendo utilizado, portanto, quando dois ou mais poços estão produzindo para uma determinada unidade. Dessa forma as vazões e pressões dos diversos poços são combinadas para a entrada na planta de processamento primário de petróleo.

Inicialmente, os fluidos passam por um separador, que pode ser bifásico, no qual ocorre a separação gás/líquido, ou trifásico, em que, além disso, também ocorre a separação óleo/água. O Quadro 2, baseado em (TRIGGIA *et al.*, 2001), demonstra os principais mecanismos envolvidos na separação líquido/gás.

Quadro 2. Mecanismos de separação utilizados na separação das fases líquida e gasosa

Mecanismo	Funcionamento
Ação da gravidade e diferença entre as densidades das fases presentes	Ocorre a decantação do fluido mais pesado
Separação inercial	São realizadas mudanças bruscas tanto na velocidade quanto na direção do fluido. Dessa forma, a fase gasosa se desprende da fase líquida devido a inércia que esta fase possui.
Aglutinação das partículas	As gotículas de óleo que estão dispersas sobre a superfície entram em contato, facilitando assim sua coalescência, aglutinação e decantação.
Força centrífuga	Utiliza-se das diferenças entre as densidades do líquido e gás para realizar a separação entre essas fases.

Fonte: Adaptado de (TRIGGIA *et al.*, 2001)

Considerando os casos nos quais forma-se uma emulsão de óleo e água, surge uma camada de água no fundo, relativamente limpa, chamada de água livre. Acima dessa camada de água existe uma camada de emulsão e na superfície de tudo, uma fina camada de óleo. Assim, os separadores trifásicos são utilizados para remoção dessa água livre presente no processo. O procedimento é muito semelhante ao dos separadores bifásicos, entretanto um espaço deve ser deixado para a decantação do líquido, bem como um dispositivo para a remoção da água (TRIGGIA *et al.*, 2001).

Essa água produzida apresenta uma variedade de compostos que podem estar dissolvidos ou dispersos. Dentre os diversos produtos químicos orgânicos existentes, ressaltam-se, principalmente, a presença dos hidrocarbonetos (NEFF; LEE; DEBLOIS, 2011). Por essa razão, essa água, antes de ser descartada precisa de passar por um tratamento adequado a fim de que seja possível recuperar parte do óleo presente nela e

em seguida, seja então, descartada ou reinjetada. Caso contrário, grandes quantidades de compostos orgânicos podem ser descartadas no mar. Tais compostos quase não são removidos quando aplicados tratamentos convencionais (KLEMZ *et al.*, 2021).

Tradicionalmente, a água produzida pelos processos de separação é encaminhada, a princípio, para um desgaseificador a fim de que traços de gás ainda presentes no líquido possam ser removidos. A partir daí, segue para um separador de água e óleo, sendo que os hidrociclones e a flotação são os processos mais comumente utilizados, uma vez que existe uma força centrífuga que direciona os componentes mais pesados (água e sólidos) contra as paredes (TRIGGIA *et al.*, 2001).

Assim, a água produzida é finalmente descartada no oceano com um determinado teor de óleos e graxas. Diferentes países propuseram diferentes métodos para a aferição do TOG (NEFF; LEE; DEBLOIS, 2011). Segundo Yang (2011), os valores de óleo em água são dependentes do método utilizado para fazer a aferição. Isso acontece porque a depender do método apenas a parte dispersa será mensurada ou ambas as partes, dispersa e dissolvida, o serão.

Nesse contexto faz-se importante entender os conceitos relacionados a óleo disperso e óleo dissolvido. O primeiro está associado ao óleo presente na água na forma de pequenas gotas, podendo variar desde submícrons até centenas de mícron. Já o segundo, refere-se ao óleo presente na água em uma forma solúvel (são os hidrocarbonetos aromáticos juntamente com ácido orgânico que formam maior parte do óleo dissolvido) (YANG, 2011).

Dentre os métodos utilizados para aferição do TOG podem-se citar os métodos infravermelho e a espectrofotometria por absorção, como métodos passíveis de serem realizados a bordo. Além do método gravimétrico que, devido às particularidades deve ser realizado em terra.

O método de absorção de infravermelho é aquele no qual uma amostra de água oleosa é acidificada e em seguida é extraída por um solvente de clorofluorcarbono. Separa-se o extrato da amostra de água, que é então seco e purificado. Parte desse extrato é colocada em um instrumento infravermelho, onde a absorbância é medida. Por fim, compara-se o valor obtido para a absorbância com os valores de outras amostras que são preparadas com concentrações conhecidas, e consequentemente a concentração de óleo na amostra original pode ser calculada (YANG, 2011).

Em relação ao método de espectrofotometria de absorção molecular, segue-se o padrão de execução PE-3UBC-02899. É importante ressaltar que esse método é aplicável à determinação de substâncias solúveis em n-hexano e que possam ser detectadas por espectrofotometria de absorção molecular no comprimento de onda de 400 nm. Extrai-se pelo menos duas vezes uma amostra que possui cerca de 500 mL com n-hexano em um funil de separação. Em seguida, o extrato é seco com sulfato de sódio, avolumado a 100 mL, em balão volumétrico e a absorbância lida em 400 nm em espectrofotômetro de absorção molecular.

Utiliza-se o comprimento de onda de 400 nm uma vez que apresenta boa sensibilidade. Uma particularidade associada ao método é que o seu limite de detecção é 0,7 mg/L e o de quantificação é de 2 mg/L, determinados experimentalmente. Além disso, não é recomendado para amostras que contenham um teor de óleos e graxas superior a 10.000 mg/L.

Os métodos baseados em gravimetria, utilizados para fins fiscais de acordo com a legislação do CONAMA ("CONAMA Resolution No. 393/2007", 2007), medem, por meio do processo de pesagem, tudo que seja passível de extração por um solvente que não seja removido durante um processo de evaporação. Nesse tipo de abordagem, tem-se, tipicamente, uma amostra de água oleosa que é extraída por um solvente. Separa-se, então, o solvente da amostra de água, sendo que o solvente passa a possuir óleo em sua composição. Em seguida, ele é colocado em um frasco cujo peso é conhecido e esse frasco é colocado em um banho de água com temperatura controlada. O solvente evapora a uma temperatura específica, e então após condensar-se ele é coletado. Por fim, o frasco contendo o óleo residual é seco e pesado e pode-se calcular a quantidade de óleo residual uma vez que o peso do frasco é conhecido (YANG, 2011).

Dado o exposto, é possível que em virtude desse descarte associado às variáveis meteo-oceanográficas, detalhadas mais a frente nesta pesquisa, surjam feições oleosas, tema principal do presente trabalho.

2.2. Efeito iridescente

O efeito iridescente é causado pela presença de determinada quantidade de óleo na água associada a incidência da luz solar. Isso pode ser um problema para plataformas de processamento primário de petróleo quando a quantidade de óleos e graxas extrapola níveis pré-estabelecidos pelos órgãos fiscalizadores e é então descartada nos oceanos.

Quando uma mancha de óleo, doravante denominada feição oleosa, é identificada e essa possui mais que 500 metros de extensão, multas ao Instituto Brasileiro do Meio Ambiente e dos Recursos Naturais (IBAMA) devem ser pagas.

O aspecto de uma feição oleosa pode variar dependendo de acordo com sua espessura, sendo que os principais fatores que determinam a espessura de uma camada de óleo é a quantidade de óleo derramado e a taxa com a qual é descartado na água (GARCIA-PINEDA *et al.*, 2020).

O acordo sobre o código de aparência do óleo de Bona (*Bonn agreement oil appearance code*) (LEWIS, 2007) destaca os principais aspectos relacionados à aparência de uma feição oleosa:

(i) Brilhante: uma camada de óleo com aspecto prateado, quase transparente, que normalmente é possível de ser detectada quando observada de um ângulo obliquo.

(ii) Arco-íris brilhante: uma camada de óleo um pouco mais espessa que a anterior que reflete as cores de um arco-íris.

(iii) Metálico: camada de espessura ainda maior que a anterior, cuja tendência é refletir a cor do céu, mas incorporando a esse reflexo alguns elementos de coloração oleosa, algo geralmente entre cinza claro e marrom opaco.

(iv) Óleo cru descontínuo: camada de óleo que reflete a verdadeira cor do óleo cuja camada apresenta espessura maior que as camadas anteriores.

(v) Óleo cru contínuo: ocorre em camadas oleosas que possuam pelo menos centenas de micrometros de espessura, apresentando continuamente uma cor escura, natural do óleo.

A Tabela 1 relaciona a aparência de uma feição oleosa com a espessura da camada de óleo derramado e a quantidade de litros de óleo por km² (LEWIS, 2007).

Tabela 1. Níveis de aparência em função da espessura da camada de óleo e da quantidade de litros por km²

Aparência	Intervalo de espessura da camada (µm)	Litros por km²
Brilhante (prateado/cinza)	0,04 a 0,30	40 – 300
Arco-íris (iridescente)	0,30 a 5,00	300 – 5.000
Metálica	5,00 a 50,00	5.000 – 50.000
Óleo cru descontínuo	50,00 a 200,00	50.000 – 200.000
Óleo cru contínuo	Superior a 200,00	Superior a 200.000

Mais recentemente, segundo a Terceira versão do *Open Water Oil Identification Job Aid for Aerial Observation* (NOAA-CODE) (NATIONAL OCEANIC AND ATMOSPHERIC ADMINISTRATION - NOAA, 2016), lançado em Agosto de 2016, as manchas de óleo podem ser classificados em 5 classes distintas que variam de acordo com a espessura da mancha. Essas classes, em ordem crescente de espessura estão elencadas a seguir:

(i) Brilhante: uma fina camada de óleo (0,005 mm), a cor pode variar de arco-íris, quando a camada é mais espessa, a prateada, ou mesmo quase transparente em camadas mais finas;

(ii) Metálico: camada mais espessa que a anterior, que tende a refletir a cor do céu, mas incorporando a esse reflexo alguns elementos de cor oleosa, geralmente entre o cinza claro e o marrom opaco;

(iii) Transição (*Transitional dark - true color*): camada de óleo que reflete a verdadeira cor do óleo cuja camada é mais espessa do que as camadas anteriores;

(iv) *Dark* (*or True*) *Color*: ocorre em camadas oleosas com pelo menos centenas de mícrons de espessura, apresentando continuamente uma cor de óleo natural e escura;

(v) Emulsão (*Emulsified oil or mousse*): é uma mistura de água em óleo que pode aparecer em diferentes tons de laranja, marrom e/ou vermelho.

A Figura 6 a seguir ilustram alguns exemplos de feições oleosas de acordo com o NOAA-CODE. As diferentes intensidades estão relacionadas a espessura da camada de óleo, conforme anteriormente apresentado.

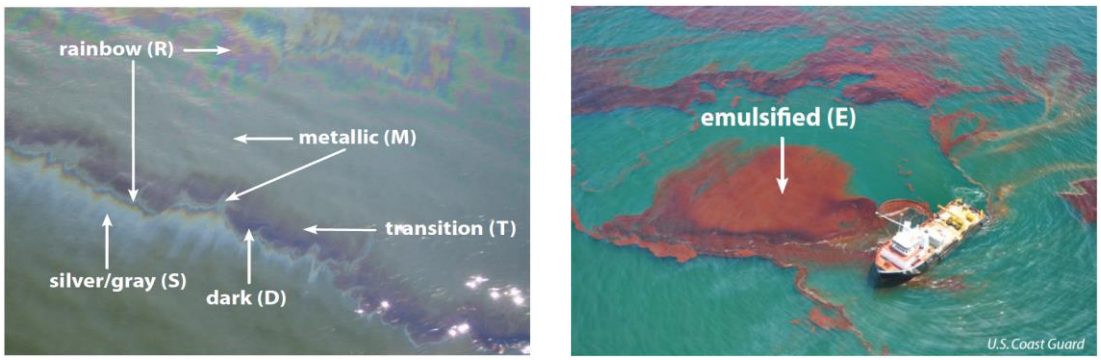


Figura 6. Exemplos de feições oleosas (NOAA-CODE) (2016)

Sabe-se ainda que variáveis meteo-oceanográficas relacionadas à corrente marítima, ao vento e às ondas, podem ter influência significativa na trajetória do óleo derramado

em água, conforme mencionado anteriormente. Dentre elas, podem-se mencionar a intensidade do vento e da corrente marítima, segundo Pisano *et al.* (2016), e as propriedades do óleo, tais como a densidade e a tensão superficial, de acordo com Zatsepa *et al.* (2018).

Tais efeitos podem apresentar um impacto perigoso e cumulativo nos ecossistemas marinhos, mesmo em pequenas quantidades. Quando aves marinhas são expostas a esse tipo de feição oleosa, é possível observar um fenômeno chamado de termorregulação reduzida, o que pode ser letal. Nesse contexto, essas aves tendem a apresentar bastante dificuldade em se recuperarem de perturbações, como por exemplo, um aumento da taxa de mortalidade em espécies adultas (O'HARA; MORANDIN, 2010).

Além da redução significativa da capacidade de termorregulação, as penas das aves marinhas quando entram em contato com o óleo, acabam perdendo a impermeabilidade característica. Assim, passa a haver penetração de água e óleo e, consequentemente, há perda da capacidade de flutuar, bem como alterações de voo. A ingestão de óleo também pode ser altamente prejudicial a tais espécies (MORANDIN; O'HARA, 2016).

Dado o exposto, verifica-se a importância de evitar o aparecimento de tais feições oleosas, uma vez que seus efeitos podem ser muito prejudiciais à biodiversidade daquela região.

2.3. Problemas de classificação

Na literatura são encontrados diversos tipos de situações que requerem a aplicação de métodos cada vez mais robustos em problemas de regressão, classificação, séries temporais e análises de cluster (DING; BAR-JOSEPH, 2017; MEHDIZADEH; FATHIAN; ADAMOWSKI, 2019; RAMAMURTHY *et al.*, 2022). Os conjuntos de dados apresentam elevada complexidade nas relações existentes entre as variáveis de entrada e as respostas investigadas, tornando o problema ainda mais complexo.

Um problema de classificação, foco da pesquisa apresentada neste documento, consiste em predizer à qual classe pertence determinado vetor de variáveis preditoras. Podem ser encontrados em diversas áreas como: processo de manufatura a fim de reconhecer o estado da taxa de desgaste da ferramenta (LI, GUOFA *et al.*, 2020), classificação de carteiras de crédito de empréstimo (SOLEYMANI; MASNAVI; SHATEYI, 2021), previsão de danos em concreto armado (THAI *et al.*, 2021), realização

do reconhecimento de modelo de espectroscopia de impedância eletroquímica (ZHU *et al.*, 2019), desenvolvimento de modelos de classificação de falhas em energia rede de distribuição do sistema (RAI; LONDHE; RAJ, 2021), e diversos outros.

Existem problemas em que um grande número de variáveis preditoras são consideradas, o que muitas vezes pode elevar a complexidade do problema. O conjunto de dados utilizado em Mansouri *et al.* (2013), por exemplo, inicialmente, possuía 41 variáveis de entrada, relacionadas às propriedades químicas e estruturais de diversos compostos e duas classes, compostos biodegradáveis e não biodegradáveis, categorizando-o como um problema de classificação binário.

Dentre os diversos métodos existentes para solução desse tipo de problema, 5 foram escolhidos para a aplicação nessa pesquisa, já que são amplamente utilizados na literatura. São eles: *Random Forest* (RF), *K-nearest neighbors* (KNN), Redes Neurais Artificiais (RNA) do tipo *Multi-Layer Perceptron* (MLP), Regressão Logística Binária (RLB) e Support Vector Machine (SVM). O Quadro 3 mostra diversas aplicações dessas técnicas de *machine leraning* em diferentes contextos. É importante ressaltar que em muitos casos descritos no Quadro 3, outras técnicas como análise de componentes principais, *partial least squares*, entre outras, são aplicadas em conjunto com as técnicas de *machine learning*.

Quadro 3. Aplicação das técnicas utilizadas abordadas em diversos estudos de diferentes áreas

Referência	Aplicação
(BALTHIS *et al.*, 2017)	Dados pareados de contaminantes de sedimentos e infaunais bentônicos foram analisados por meio de regressão logística para derivar referências de qualidade de sedimentos a fim de avaliar riscos de impactos relacionados ao petróleo.
(BRANDÃO; BRAGA; SUAREZ, 2012)	Determinação de adulterantes de óleos e gorduras vegetais no óleo diesel aplicando técnicas multivariadas e KNN
(CHEN, YIFU *et al.*, 2021)	Utilização de diversas técnicas como RF, RLB, SVM, KNN, entre outros, para análises de *oil fingerprints* em cenários de derramamento de óleo.
(CUI *et al.*, 2015)	Classificação de compostos seletivos do receptor de estrogênio por meio de RNA.
(FU *et al.*, 2017)	Utilização de SVM para classificação e triagem de biomarcadores em metabolômica.
(GHORBANZAD'E; FATEMI, 2012)	Classificação dos agentes do sistema nervoso central utilizando SVM.

Continuação do Quadro 3

Referência	Aplicação
(HE *et al.*, 2021)	Utilização do RF na previsão do desempenho do calor de detonação de compostos contendo nitrogênio com base em QSAR e RF.
(HEGDE; MILLWATER; GRAY, 2019)	Algoritmos como RLB, SVM e RF foram utilizados para classificação da severidade do stick-slip de perfuração.
(ISLAM *et al.*, 2014)	Gerenciamento de resíduos sólidos aplicando RNA.
(JIMÉNEZ-CORDERO; MORALES; PINEDA, 2021)	Nova abordagem para seleção de características para classificação utilizando SVM. Os testes foram aplicados em conjuntos de dados de benchmark.
(LI, HONGDONG; LIANG; XU, 2009)	Aborda diversas aplicações de SVM na área química.
(LIANG *et al.*, 2020)	RF é aplicado no monitoramento da qualidade de um importante medicamento chinês (*salvia miltiorrhiza*).
(LIN, SHUN KU *et al.*, 2021)	Classificação de pacientes com Alzheimer por meio da utilização de RNA.
(LIU, FANG; ZHOU, 2015)	Nova metodologia combinando SVM com algoritmos de otimização (*particle swarm*). Os testes foram realizados em diversos banco de dados, inclusive o da flor íris.
(MARINS *et al.*, 2021)	RF é utilizado para detectar e classificar falhas em poços de petróleo e linhas de produção.
(MORAIS; LIMA, 2017)	Classificação de dados de matriz de emissão de excitação fluorescente utilizando SVM.
(OLOSO *et al.*, 2018)	*Ensemble* de SVM para avaliação da viscosidade de petróleo cru.
(OTCHERE *et al.*, 2021)	Predição das propriedades de reservatórios de petróleo utilizando técnicas como RNA e SVM.
(PERES *et al.*, 2011)	Utilização RNA para classificação de frutos por cultivar de oliva, a fim de garantir a autenticidade varietal.
(QUINTANILHA *et al.*, 2021)	Detecção e classificação dos vários tipos de falhas em um sistema de geração de energia aplicando RF.
(SARAVANA KUMAR; MANIKANDAN, 2018)	Classificação de *big data* da área de medicina (prontuário do paciente, detalhes do medicamento e dados da equipe, etc.) aplicando RF.
(SAVOLAINEN; KAZMIERCZAK; CAFFEY, 2013)	Regressão logística foi utilizada para determinar o efeito de acidentes ambientais em uma pesquisa da indústria de pesca recreativa de aluguel do Golfo do México.
(SPEISER *et al.*, 2019)	Utilização do RF, em conjunto com outras técnicas, a fim de realizar a previsão de resultados binários clusterizados e conjuntos de dados com muitas variáveis preditoras.
(YEAP *et al.*, 2020)	Utilização do RF em dados de cromatografia gasosa / espectrometria de mobilidade diferencial.

Continuação do Quadro 3

Referência	Aplicação
(ZADKARAMI; SHAHBAZIAN; SALAHSHOOR, 2016)	Detecção da ocorrência de falha de vazamento em oleodutos de hidrocarbonetos, além de sugerir sua localização e severidade utilizando RNA.
(ZHANG, TIANLONG *et al.*, 2016)	Classificação e reconhecimento de nove graus de aço aplicando RF.

Fonte: O autor

Um conceito importante quando se trabalha com problemas de classificação é a chamada matriz de confusão. É uma das formas mais compreensíveis de se representar os resultados de uma classificação binária. Nesse caso, obtém-se uma matriz com duas linhas e duas colunas, sendo que na diagonal principal localizam-se a quantidade de acertos e na diagonal secundária as classificações incorretas (MÜLLER; GUIDO, 2016). Um exemplo de uma matriz de confusão para um problema binário é apresentado na Figura 7.

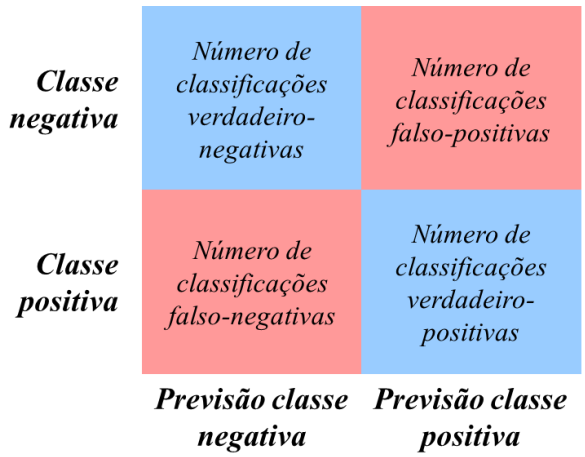


Figura 7. Exemplo de matriz de confusão (adaptado de (MÜLLER; GUIDO, 2016))

Métricas comumente utilizadas nesse tipo de problema são a acurácia (A_c), a especificidade (S_p) e sensibilidade (S_n) para avaliar o desempenho do classificador no conjunto de dados de teste. De acordo com Wan *et al.* (2018) e Mansouri *et al.* (2013), os valores de A_c, S_p e S_n podem ser calculados conforme mostrado nas equações Eq. (1), Eq. (2) e na Eq. (3), respectivamente, sendo que *TN*, *FN*, *TP* e *FP* indicam o número de verdadeiros negativos, falsos negativos, verdadeiros positivos e falsos positivos, respectivamente.

$$A_c = \frac{TP + TN}{TN + FP + TN + FN} \quad (1)$$

$$S_p = \frac{TN}{TN + FP} \quad (2)$$

$$S_n = \frac{TP}{TP + FN} \quad (3)$$

Outra estratégia comumente utilizada em pesquisas que envolvem problemas de classificação é a obtenção da Característica de Operação do Receptor (*Receiver Operation Caracteristic*), também conhecida como curva ROC (HEGDE; MILLWATER; GRAY, 2019). Em muitos casos, o *threshold* utilizado em um problema de classificação binário é 50%, isso significa que uma probabilidade de pertencimento a classe 1 superior a esse valor, resultará na classificação da observação como classe 1 e o contrário resultará na classificação da observação como pertencente à classe 0. Entretanto esse limiar pode variar de acordo com o desejo do pesquisador.

Após o preenchimento da matriz de confusão, é possível calcular as taxas de classificações verdadeiro-positivas (*True Positive Rate* – TPR) e a taxa de classificações falso-positivas (*False Positive Rate* – FPR). É importante ressaltar que o valor da TPR é a própria especificidade, enquanto que o valor da FPR é igual a $1 - S_n$ (FAWCETT, 2006).

Assim, após treinar classificador selecionado, uma matriz de confusão distinta é obtida à medida que o valor do *threshold* varia. Dessa forma, é possível desenhar uma curva semelhante àquela representada na Figura 8. Convém destacar que quanto maior a área sob a curva ROC, melhor o desempenho do classificador.

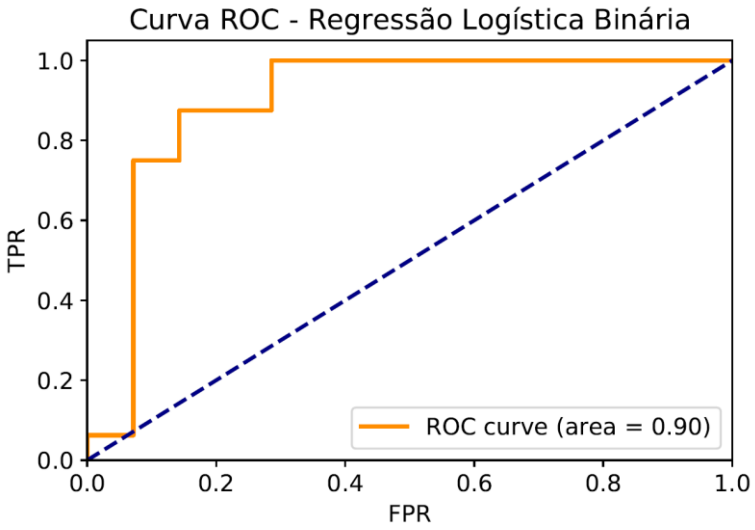


Figura 8. Exemplo de curva ROC para regressão logística binária

Conforme pode ser observado na Figura 8, o ponto $(1,1)$ indica que o *threshold* é tão baixo, que as probabilidades de pertencimento à classe positiva para todas as observações são superiores a ele. Assim, a TPR é igual a 1 uma vez que todos os pontos da amostra foram classificados como pertencentes à classe positiva. Por outro lado, todas as observações que pertenciam à classe negativa foram erroneamente classificadas, levando a uma FPR também igual a 1. Algo similar acontece com o ponto $(0,0)$, uma vez que o valor do *threshold* é tão alto que todas as observações são classificadas como pertencentes à classe negativa. Dessa forma chega-se a uma FPR igual a 0, entretanto a uma TPR também igual a zero.

A linha tracejada mostrada no gráfico da Figura 8 indica os pontos onde TPR = FPR, ou seja, a taxa de observações corretamente classificadas como pertencentes à classe positiva é a mesma daquelas classificadas erroneamente classificadas como positivas. A fim de que um classificador seja considerado adequado, é necessário que no mínimo a área sob a curva seja superior à área sob essa linha tracejada. Maiores explicações sobre esse gráfico, bem como os métodos para os cálculos da área sob a curva podem ser encontrados em (FAWCETT, 2006).

2.4. *Random Forest*

Uma árvore de decisão pode ser entendida como uma técnica analítica de dados que aprende hierarquicamente a partir de uma série de questões do tipo *if*/*else*. A partir de tais questionamentos chega-se a uma decisão final, podendo ser relacionada tanto a um problema de regressão quanto de classificação. No caso de dados contínuos, frequentemente encontrados na literatura, as questões são elaboradas de modo a definir se determinada variável de entrada é maior ou igual ($\geq$) ou menor ou igual ($\leq$) a um valor a previamente especificado (MÜLLER; GUIDO, 2016).

Assim, por meio do uso das árvores de decisão é possível explorar complexas interações presentes no conjunto de dados (YONG et al., 2020). A Figura 9 apresenta um esquema de uma árvore de decisão para um conjunto de dados apresentado em Müller e Guido (2016). Esse conjunto possui duas variáveis de entrada x_1 e x_2 e 100 observações representadas por triângulos vermelhos (*V*) e círculos azuis (*A*).

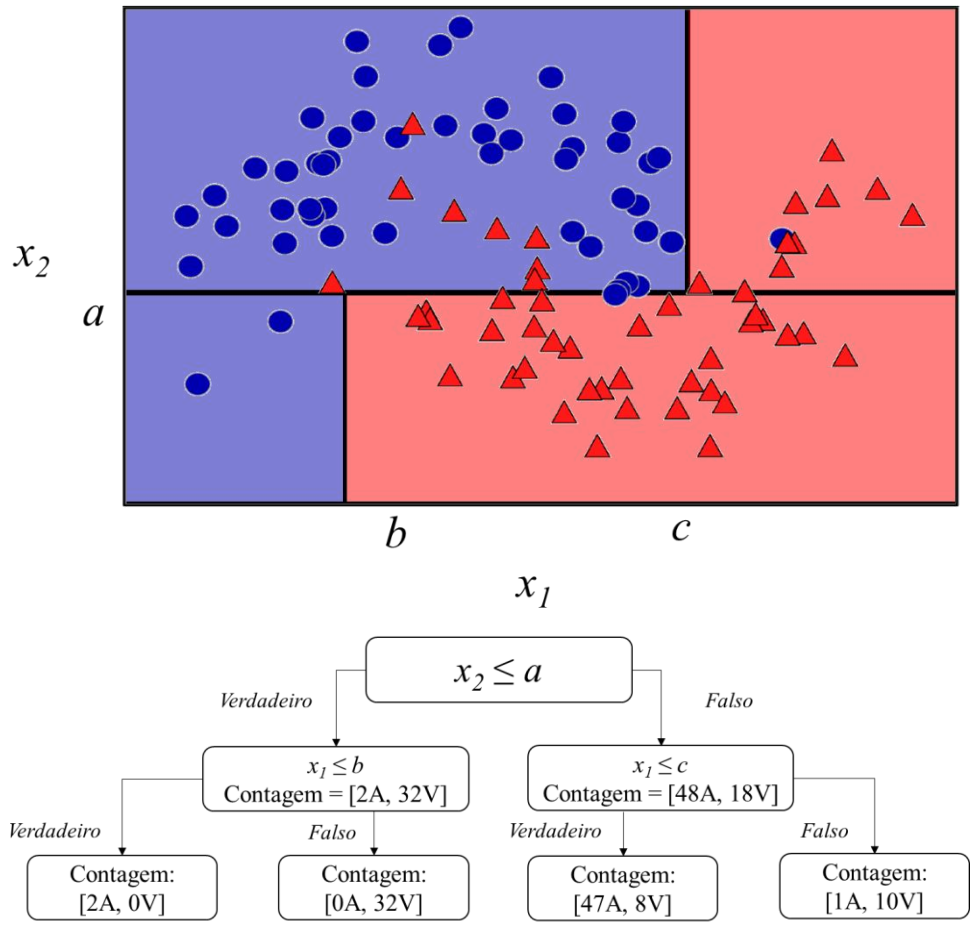


Figura 9. Árvore de decisão ilustrativa adaptada de (MÜLLER; GUIDO, 2016)

A partir disso, o método denominado Floresta Randômica (*Random Forest* – RF) pode ser definido como uma combinação de muitas árvores de decisão em que as árvores são significativamente diferentes umas das outras. Com ele, espera-se melhorar a capacidade de generalização de uma única árvore (MÜLLER; GUIDO, 2016). Além disso, não requer uma especificação prévia de relações não lineares entre as variáveis e também pode fornecer informações sobre a importância dos preditores durante o processo de classificação (SPEISER *et al.*, 2019). Nesse sentido, pode-se afirmar que a RF compartilha todas as vantagens das árvores de decisão, mas também pode superar muitas de suas deficiências (MÜLLER; GUIDO, 2016).

Muitos estudos já demonstraram que o método RF apresenta um processo de aprendizado rápido, melhor tolerância ao ruído, e é capaz de fornecer a importância associada às variáveis de entrada, conforme mencionado anteriormente (HE *et al.*, 2021). Dois parâmetros são realmente importantes para este algoritmo, que são o número de árvores de decisão e o número de atributos selecionados aleatoriamente (geralmente, quanto menor, menor a chance de *overfitting*) (HE *et al.*, 2021; MÜLLER; GUIDO, 2016).

É importante destacar que caso o número de árvores de decisão aumente consideravelmente, ainda assim isso não acarretará *overfitting*, o que pode ser explicado pela lei dos grandes números, conforme elucidado por Breiman (2001).

No entanto, é importante ressaltar que RF não apresenta um excelente desempenho em lidar com dados dimensionais muito elevados ou dados esparsos, ou seja, matrizes que contém uma grande quantidade de posições cujo valor é zero. Além disso, pode exigir mais memória e tempo quando comparado aos modelos lineares. Nestes casos, cabe ao pesquisador resolver esse *trade-off* (MÜLLER; GUIDO, 2016).

Um importante resultado obtido por meio da aplicação dessa técnica é o gráfico de importância das características (*feature importance plot*), ou seja, esse gráfico apresenta a importância associada a cada uma das variáveis de entrada. Um alto valor associado a determinada variável indica que ela é importante para a tomada de decisão, enquanto que um valor próximo de zero indica que esta variável praticamente não é levada em consideração para a decisão final, que por exemplo, pode ser previsão do valor de uma determinada variável resposta (DAI *et al.*, 2018). Convém destacar que a soma dessas importâncias sempre resulta em 1.

As etapas do algoritmo RF, de maneira sintetizada e considerando um problema de classificação, podem ser visualizadas na Figura 10 baseada em (MUSBAH; ALY; LITTLE, 2021).

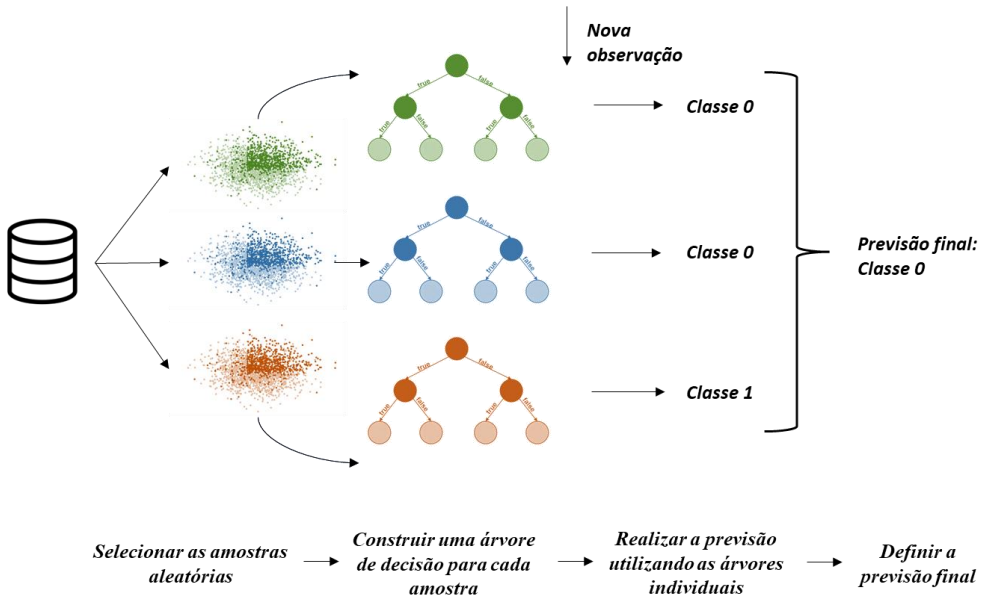


Figura 10. Etapas do método Random Forest

Inicialmente são selecionadas diversas amostras do conjunto de dados de forma aleatória. A seguir uma árvore de decisão é construída para cada uma das amostras, sendo que ao chegar uma nova observação cada árvore irá classifica-la em uma determinada classe, sendo que a previsão final é definida com base nas classificações das árvores.

2.5. K-vizinhos mais próximos (*K-nearest neighbors*)

O algoritmo conhecido como K-vizinhos mais próximos (KNN - *K-nearest-neighbors*) tem sido amplamente utilizado em problemas de reconhecimento de padrões (ABU-AISHEH; RAVEAUX; RAMEL, 2020). Trata-se de um método não paramétrico, baseado em instâncias e é considerado um dos mais simples e também mais importantes dentre os diversos algoritmos de *machine learning* (SANQUETTA *et al.*, 2018; ZHANG, SHICHAO; CHENG; *et al.*, 2018; ZUO; ZHANG; WANG, 2008).

O algoritmo baseia-se em armazenar os valores do conjunto de treinamento e realizar cálculos de distâncias sempre que nova previsão for realizada (MÜLLER; GUIDO, 2016). Isso o diferencia dos demais algoritmos, uma vez que métodos como RF, MLP, SVM e RL constroem um modelo a partir do banco de dados fornecidos e utilizam-se do modelo construído para realizar novas previsões. Sendo assim, para esses últimos algoritmos, o conjunto de dados de treinamento só se torna necessário caso seja de interesse do pesquisador 'recalibrar' o modelo.

A proposta do KNN, conforme indica o próprio nome, é encontrar os *k* vetores de dados mais próximos ao novo vetor de dados a ser classificado. Dessa forma, a nova previsão será a classe mais frequente entre os *k* vizinhos mais próximos (ABU-AISHEH; RAVEAUX; RAMEL, 2020; MÜLLER; GUIDO, 2016).

O método é de fácil compreensão e apresenta formulas conhecidas para os cálculos. Os valores das distâncias, por exemplo, são frequentemente obtidos considerando a distância Euclidiana (BASSIOUNI *et al.*, 2018; EL-DAHSHAN; BASSIOUNI, 2018; PALANISAMY; MURUGAPPAN; YAACOB, 2013; YELIPE; PORIKA; GOLLA, 2018). A Eq. (4) exemplifica o cálculo dessa distância de acordo com El-Dahshan e Bassiouni (2018), sendo que $\boldsymbol{X}$ e $\boldsymbol{Y}$ representam dois vetores com um total de p componentes.

$$d(\boldsymbol{X}, \boldsymbol{Y}) = \sum_{k=1}^{p} \|X_k - Y_k\|^2 \qquad (4)$$

Outra distância comumente utilizada é a de Minkowski, conforme mostrado na Eq. (5), sendo q um valor não negativo denominado coeficiente de Minkowski. A partir desta distância, também pode-se obter a distância de Manhattan, bastando assumir $q = 1$ (ZHANG, SHICHAO, 2012).

$$d(\boldsymbol{X}, \boldsymbol{Y}) = \left(\sum_{k=1}^{p} |x_k - y_k|^q \right)^{1/q} \tag{5}$$

Geralmente, a distância Euclidiana funciona adequadamente para a maioria dos conjuntos de dados (MÜLLER; GUIDO, 2016). Entretanto, uma desvantagem do método é que para um conjunto de dados muito grande pode consumir muito tempo e esforço computacional (ABU-AISHEH; RAVEAUX; RAMEL, 2020). Além disso, é importante que os dados estejam normalizados ou padronizados, pois pode ser dada, erroneamente, maior importância às variáveis de maior grandeza, conforme pode ser facilmente inferido por meio das equações demonstradas anteriormente.

Finalmente, a escolha do parâmetro k é de extrema importância para um melhor desempenho do método. Segundo Hassanat *et al.* (2014), na maioria dos casos esse parâmetro é escolhido empiricamente, variando de acordo com o conjunto de dados com o qual se trabalha. Por outro lado, muitos trabalhos são encontrados na literatura mostrando diferentes técnicas a serem utilizadas a fim de encontrar um valor ótimo para esse parâmetro (HASSANAT *et al.*, 2014; ZHANG, SHICHAO; LI; *et al.*, 2018).

Uma síntese do método KNN aplicado em um conjunto de classificação pode ser observada na Figura 11 com base em (MÜLLER; GUIDO, 2016).

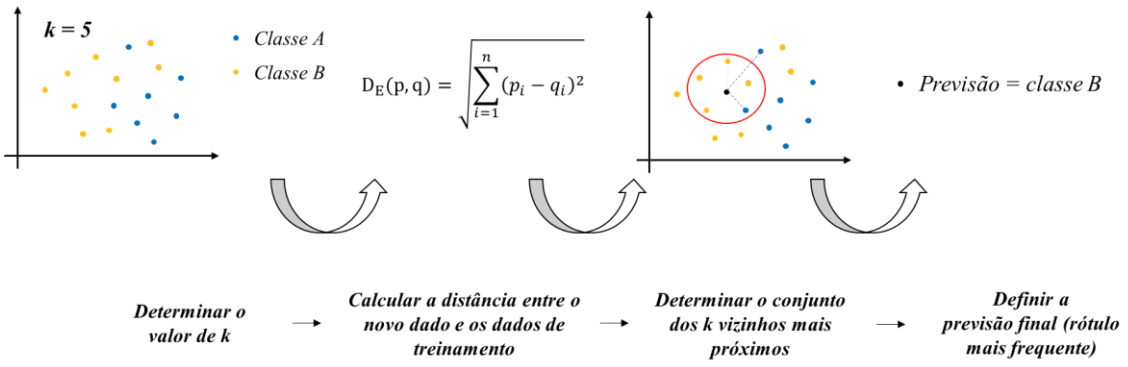


Figura 11. Etapas do algoritmo KNN

2.6. Redes neurais artificiais

Redes Neurais Artificiais (RNA) é outra abordagem comumente usada que é capaz de entender relações complexas entre as variáveis de entrada e saída em um determinado conjunto de dados. O uso de RNA pode ser frequentemente visto em problemas de classificação (GANBOLD; CHASIA, 2017; ISLAM *et al.*, 2014), análise de cluster (OLSON; VALOVA; MICHEL, 2016; PUGGINA BIANCHESI *et al.*, 2019), regressão (KUO; FARICHA, 2016; YILMAZ; KAYNAR, 2011) e séries temporais (AIZENBERG *et al.*, 2016; BALESTRASSI *et al.*, 2009).

De acordo com Silva *et al.* (2017), algumas características importantes relacionadas à aplicação de redes neurais são a habilidade de aprender pela experiência e a capacidade de generalização. O conhecimento é armazenado em cada uma das muitas sinapses existentes, e as arquiteturas de RNA podem ser, em grande parte dos casos, prototipados em hardware ou software, já que a saída geralmente é obtida por meio de operações matemáticas.

Entre as arquiteturas *feedforward* mais conhecidas, o Perceptron multicamadas (*Multilayer Perceptron* - MLP) pode ser considerado como um dos mais citados na literatura, apresentando diversas aplicações práticas (CHAU; TRAN; DAO, 2021; ISLAM *et al.*, 2014; LIN, SHUN KU *et al.*, 2021; PERES *et al.*, 2011; WANG; MOAYEDI; KOK FOONG, 2020). O MLP consiste em uma camada de entrada, uma ou mais camadas ocultas e uma camada de saída (GANBOLD; CHASIA, 2017). É importante destacar que a complexidade da rede neural é proporcional ao número de neurônios em suas camadas. Quanto maior o número de neurônios, mais complexa é a rede, no entanto, a probabilidade de *overfitting* também aumentará. Além disso, diversos outros parâmetros devem ser considerados na parametrização da RNA, conforme amplamente discutido em (BALESTRASSI *et al.*, 2009).

A retropropagação (*backpropagation*) é comumente usada como algoritmo de aprendizado em se tratando do estágio de treinamento de uma rede do tipo MLP e pode ser dividida em duas etapas principais. O primeiro é chamado de propagação direta, uma vez que os sinais advindos do conjunto de dados se propagam pelas camadas e uma saída é produzida. Por fim, a resposta produzida é comparada com as respostas reais disponíveis no conjunto de treinamento, dessa forma os erros calculados serão utilizados para ajustar os pesos da rede (SILVA *et al.*, 2017). Vale ressaltar que se trata de um processo de

aprendizagem supervisionado, ou seja, a rede terá acesso aos valores reais durante a fase de treinamento, adquirindo a capacidade de generalizar o conhecimento e fazer previsões considerando novos dados de entrada.

Uma síntese do algoritmo utilizado por uma RNA do tipo MLP, *feedforward* e treinada segundo o algoritmo *backpropagation*, pode ser observado na Figura 12.

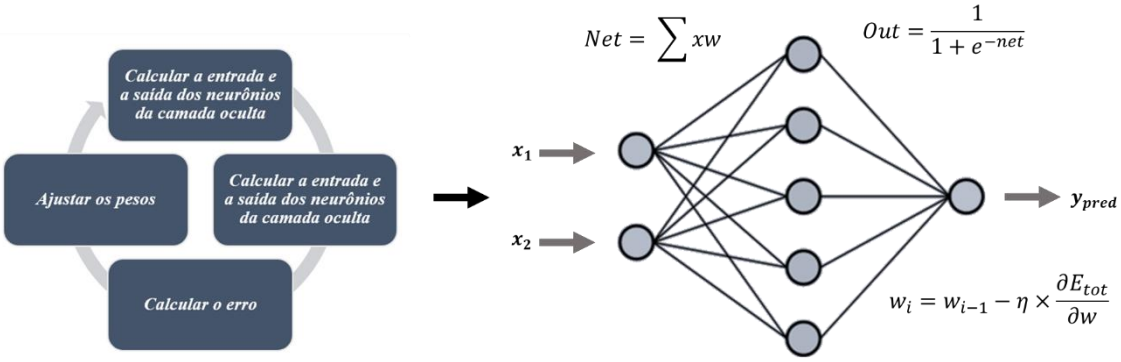


Figura 12. Etapas do algoritmo de backpropagation (adaptado de (SILVA *et al.*, 2017))

Uma explicação mais detalhada sobre RNA e as inúmeras arquiteturas existentes pode ser encontrada em (HAYKIN, 2009).

2.7. Regressão logística binária

A regressão logística binária pode ser entendida como uma técnica estatística para prever a probabilidade de um conjunto de variáveis preditoras pertencerem a determinada classe, por exemplo, 0 ou 1 (BISSACOT *et al.*, 2016). Assim como no caso da regressão linear, um modelo de regressão logística é capaz de lidar com diversas variáveis mesmo estando em escalas distintas (HOSMER JR.; LEMESHOW; STURDIVANT, 2013).

Considerando que a probabilidade de determinada variável resposta Y pertencer à classe 1, dado um vetor de variáveis de entrada composto de p elementos, pode ser representada por $P[Y = 1|\boldsymbol{x}] = \pi(\boldsymbol{x})$. Assumindo-se ainda a função *logit* para esse problema, conforme mostrado na Eq. (6), é possível estimar a probabilidade $\pi(\boldsymbol{x})$ por meio da Eq. (7) (HOSMER JR.; LEMESHOW; STURDIVANT, 2013).

$$g(x) = \ln\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p \tag{6}$$

$$\pi(x) = \frac{e^{g(x)}}{1+e^{g(x)}} \tag{7}$$

A função sigmoide, que é simplesmente a inversa da função *logit*, também é comumente citada e pode ser vista na Eq. (8). Dessa forma, a probabilidade poderá ser encontrada por meio da Eq. (9) (BISSACOT *et al.*, 2016).

$$G(t) = \frac{1}{1 + e^{t}} \tag{8}$$

$$P[c|\boldsymbol{X_t}] = G(\hat{\beta}_0 + X^T \widehat{\boldsymbol{B}}) \tag{9}$$

A Figura 13 demonstra as etapas para o algoritmo de regressão logística considerando a utilização da função *logit* mencionada anteriormente.

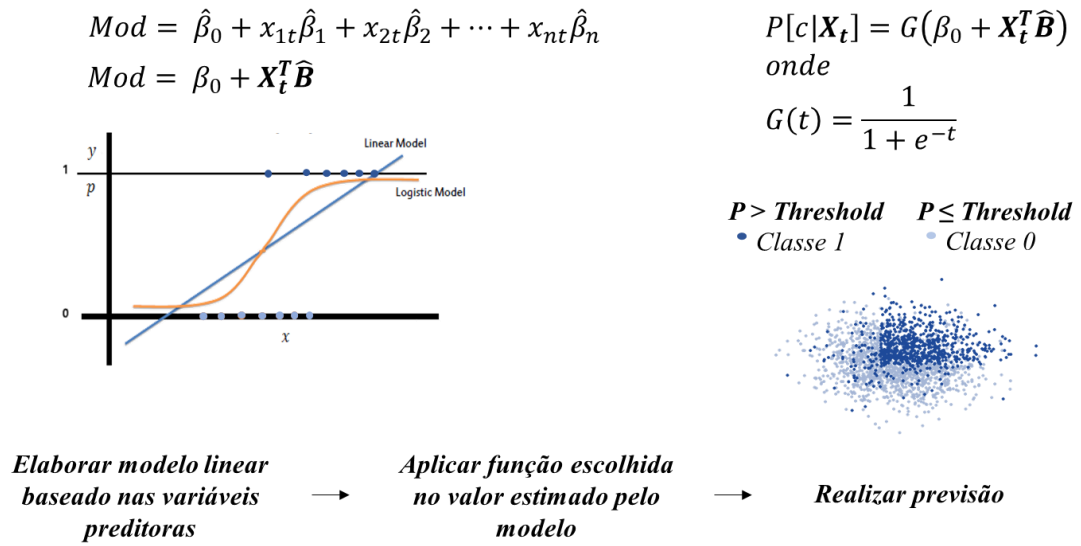


Figura 13. Etapas do algoritmo de regressão logística binária considerando a função *logit*

2.8. Máquina de vetor de suporte (*Support vector machine*)

O método SVM com kernel, geralmente referido como SVM, é outro tipo de modelo supervisionado usado em problemas de classificação. O SVM, durante a fase de treinamento, aprende a importância de cada observação do treinamento para definir o limite que separa as distintas classes. Normalmente, apenas alguns pontos são necessários para este trabalho, ou seja, aqueles pontos que ficam na fronteira que separa as classes, também chamados de vetores de suporte (MÜLLER; GUIDO, 2016).

A função Kernel aplicada pelo método SVM pode ser vista como uma técnica de aumento da dimensionalidade que permite separar linearmente dados que antes eram linearmente inseparáveis (LI, HONGDONG; LIANG; XU, 2009). Quando se trata de prever um novo ponto, é necessário calcular a distância até os vetores de suporte

considerando a importância de cada um deles, o que foi aprendido durante a fase de treinamento (MÜLLER; GUIDO, 2016).

O método SVM são capazes de lidar com um conjunto de dados contendo apenas alguns ou até mesmo muitos recursos. Em relação ao número de observações, pode fornecer resultados satisfatórios trabalhando com 10.000 amostras, por exemplo, porém trabalhar com 100.000 amostras ou mais, pode representar um desafio em termos de tempo e memória (MÜLLER; GUIDO, 2016).

Os principais parâmetros são a regularização C, o kernel escolhido e seus parâmetros específicos (MÜLLER; GUIDO, 2016). É importante destacar que o hiperplano, capaz de separar as classes investigadas, é construído por meio de uma função *kernel*, que pode ser linear, sigmoidal, polinomial ou *Radial Basis Function* (RBF) (MITICHE *et al.*, 2018). O valor de *C* pode ser entendido como o parâmetro que penaliza cada classificação errada feita pelo modelo, e disso, pode-se inferir que menores valores desse parâmetro podem levar a uma melhor capacidade de generalização (ERIŞTI; UÇAR; DEMIR, 2010). Finalmente, o valor de gama determina até qual ponto irá influenciar a linha de separação entre as classes, já que especifica (YU, XINLIANG; YU; ZENG, 2014).

A Figura 14, adaptada de Müller e Guido (2016) demonstra as etapas principais do algoritmo de SVM usados em problemas de classificação.

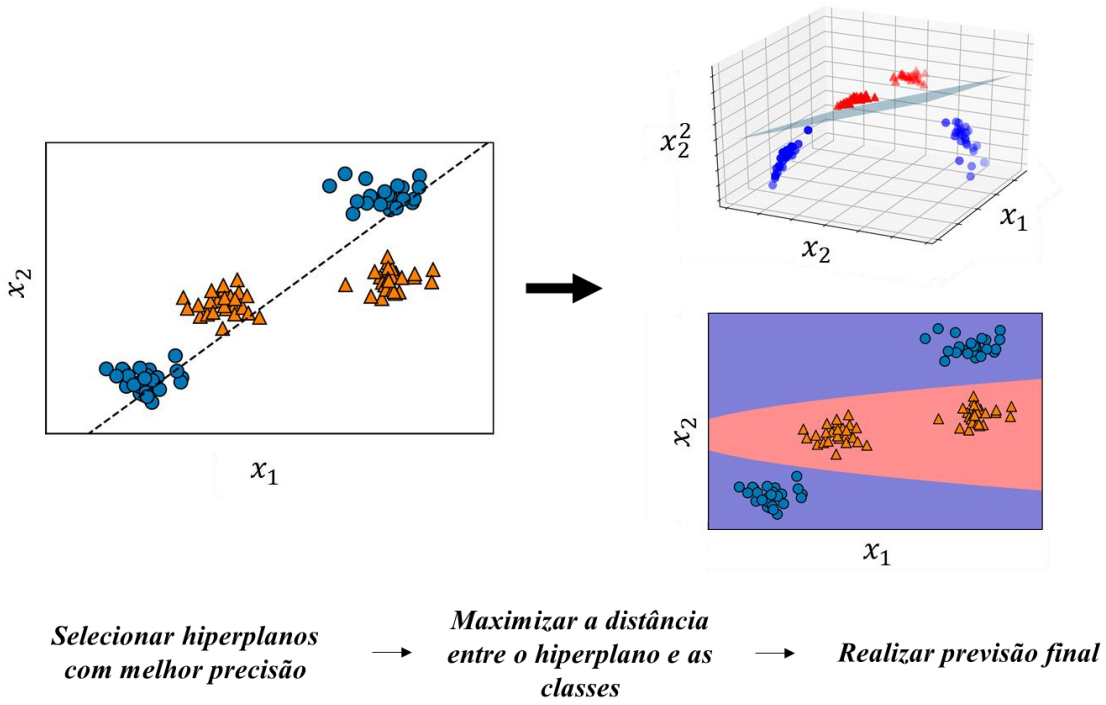


Figura 14. Etapas de do algoritmo SVC (adaptado de (MÜLLER; GUIDO, 2016))

Inicialmente deve-se selecionar hiperplanos que apresentem melhores valores de acurácia. Em seguida a distância entre os hiperplanos e as classes devem ser maximizadas. Podem ser aplicados diversos tipos de *kernels* para melhorar a precisão do modelo final. No caso da figura apresentada, acrescentou-se um outro termo, x_2^2 o que claramente facilitou a distinção entre as duas classes apresentadas.

2.9. Metodologia de superfície de resposta

A Metodologia de Superfície de Resposta (*Response surface Methodology* – RSM) é uma metodologia de extrema importância que pode ser usada para modelar e analisar um grande número de problemas. Sendo o RSM um tipo de planejamento de experimentos (*Design of Experiments* – DOE), a investigação do problema geralmente alcança bons resultados a partir de um pequeno grupo de execuções (MONTGOMERY, 2017; MYERS; MONTGOMERY; ANDERSON-COOK, 2016).

Embora amplamente conhecida no contexto de processos industriais, a literatura também apresenta uma grande adequação do RSM em diversos processos (BELINATO *et al.*, 2019; HAMEED *et al.*, 2021; MANGILI *et al.*, 2015; RIBEIRO *et al.*, 2010; ROCHA *et al.*, 2020; SACHANIYA *et al.*, 2020; SALEM; JAFARZADEH-GHOUSHCHI, 2016; VIACAVA; ROURA; AGÜERO, 2015; YU, KEQIANG *et al.*, 2019).

O primeiro passo da metodologia de superfície de resposta consiste em encontrar uma aproximação adequada para a relação existente entre a variável dependente y e o conjunto de variáveis independentes x. Um modelo de baixo grau pode ser utilizado como um polinômio de primeira ordem. Entretanto, caso exista curvatura, um polinômio de maior grau, como por exemplo de segundo grau, deve ser utilizado (MONTGOMERY, 2017). Geralmente, um modelo de segunda ordem, cuja formulação genérica é apresentada na Eq. (10), é suficiente para representar matematicamente diversas relações entre as variáveis resposta e preditoras (SACHANIYA *et al.*, 2020; SALEM; JAFARZADEH-GHOUSHCHI, 2016; VIACAVA; ROURA; AGÜERO, 2015; YU, KEQIANG *et al.*, 2019).

$$Y = \beta_0 + \sum_{i=1}^{l} \beta_i x_i + \sum_{i=1}^{l} \beta_{ii} x_i^2 + \sum_{i<j} \sum \beta_{ij} x_i x_j + \varepsilon \qquad (10)$$

Muito provavelmente o modelo de segundo grau não será adequado para representar a relação existente entre a resposta e as preditoras em todo o espaço possível das variáveis

independentes, entretanto para uma pequena região, tais modelos funcionam satisfatoriamente (MONTGOMERY, 2017).

Assim, é importante ressaltar que a metodologia de superfície de resposta é um procedimento sequencial. Comumente não se parte inicialmente da região de ótimo, dessa forma o modelo de primeira ordem é adequado. Uma vez que existe uma curvatura no sistema, é necessário buscar a região de ótimo. Montgomery (2017) apresenta o método da subida mais íngreme como forma de chegar à região de ótimo. A Figura 15 apresenta um esquema representativo da natureza sequencial da metodologia de superfície de resposta.

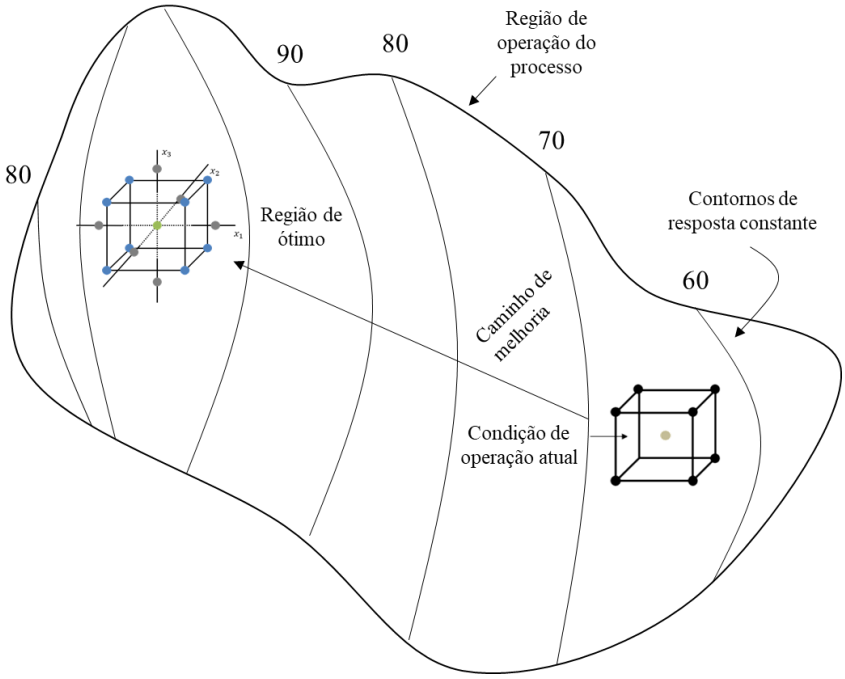


Figura 15. Representação da natureza sequencial da metodologia de superfície de resposta adaptada de (MONTGOMERY, 2017)

Um arranjo comumente utilizado para ajustar um modelo de segunda ordem é o tipo *Central Composite Design* (CCD) consiste em um arranjo com 2^k pontos fatoriais, n_c pontos centrais e 2^k pontos axiais (MONTGOMERY, 2017). No presente trabalho, a análise não partiu da experimentação, mas de um conjunto de dados já existente a partir do qual as variáveis codificadas originais foram convertidas em fatores e, portanto, um CCD customizado foi criado a partir dos escores dessas variáveis latentes.

Através da análise deste CCD, um modelo matemático em função dos fatores utilizados pode ser alcançado. A Figura 16 mostra a estrutura dos experimentos em um arranjo do tipo CCD.

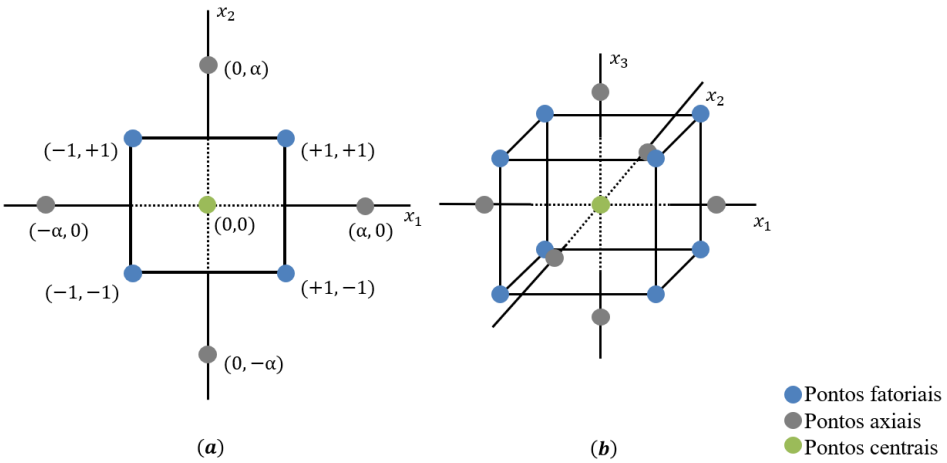


Figura 16. Representação esquemática de um CCD com 2 fatores (a) e com 3 fatores (b) adaptada de (MONTGOMERY, 2017)

2.10. Análise fatorial

Em problemas reais, frequentemente diversas variáveis de entrada são consideradas e mais de uma única resposta são investigadas. Além disso, em muitos casos, as variáveis são correlacionadas e os procedimentos tradicionais, como regressão univariada e testes estatísticos, podem não ser adequados nesses contextos. Diante disso, algumas técnicas multivariadas devem ser aplicadas a fim de evitar conclusões errôneas sobre os dados.

Segundo Johnson e Wichern (2007), a análise fatorial pode ser considerada uma extensão da análise de componentes principais (*Principal Component Analysis* – PCA). Na análise fatorial, as variáveis originais são escritas em função dos fatores não correlacionados entre si, mas altamente correlacionados com as variáveis originais.

Considerando um vetor aleatório $\boldsymbol{X}$ $(p \times 1)$ que consiste em p variáveis observáveis, com vetor de médias $\boldsymbol{\mu}$ $(p \times 1)$ e matriz de covariância $\boldsymbol{\Sigma}$ $(p \times p)$, então o modelo de fator ortogonal com m fatores comuns (onde $m < p$) pode ser escrito conforme mostrado na Eq. (11) (JOHNSON; WICHERN, 2007). $\boldsymbol{L}$ $(p \times m)$ representa os carregamentos (*loadings*), ou seja, os valores de correlação entre as variáveis originais e os fatores comuns, $\boldsymbol{F}$ $(m \times 1)$ representa os fatores comuns e $\boldsymbol{\varepsilon}$ $(p \times 1)$ é o vetor de erros associado.

$$\boldsymbol{X} = \boldsymbol{\mu} + \boldsymbol{LF} + \boldsymbol{\varepsilon} \tag{11}$$

Algumas suposições devem ser feitas, considerando o modelo ortogonal mostrado na Eq. (11). Portanto, essas suposições apresentadas na Eq. (12) juntamente com a Eq. (11) completam os conceitos associados ao modelo ortogonal (JOHNSON; WICHERN, 2007).

$$E(\boldsymbol{F}), \qquad Cov(\boldsymbol{F}) = E[\boldsymbol{F}\boldsymbol{F}'] = \boldsymbol{I}$$

$$E(\boldsymbol{\varepsilon}) = 0, \qquad Cov(\boldsymbol{\varepsilon}) = E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'] = \boldsymbol{\Psi} = \begin{bmatrix} \psi_1 & 0 & \cdots & 0 \\ 0 & \psi_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \psi_p \end{bmatrix} \tag{12}$$

$$Cov(\boldsymbol{\varepsilon}, \boldsymbol{F}) = E(\boldsymbol{\varepsilon}\boldsymbol{F}') = 0$$

Este modelo, de acordo com Johnson e Wichern (2007), implica nas deduções mostradas na Eq. (13).

$$\begin{aligned} \boldsymbol{\Sigma} &= Cov(\boldsymbol{X}) = E(\boldsymbol{X} - \boldsymbol{\mu})(\boldsymbol{X} - \boldsymbol{\mu})' \\ &= \boldsymbol{L}E(\boldsymbol{F}\boldsymbol{F}')\boldsymbol{L}' + E(\boldsymbol{\varepsilon}\boldsymbol{F}')\boldsymbol{L}' + \boldsymbol{L}E(\boldsymbol{F}\boldsymbol{\varepsilon}') + E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}') \\ &= \boldsymbol{L}\boldsymbol{L}' + \boldsymbol{\Psi} \end{aligned} \tag{13}$$

Aplicando a análise fatorial, as variáveis correlacionadas originais serão representadas pelos *scores* não correlacionadas dos fatores comuns. Dessa forma, os métodos estatísticos tradicionais ainda poderão ser aplicados.

Um procedimento comum realizado durante a análise fatorial é a rotação dos eixos, promovendo a aproximação dos fatores dos *loadings* dos fatores. Esse procedimento é realizado antes da extração dos *scores* dos fatores e gera modelos mais simples e de fácil compreensão baseado no princípio da parcimônia (THURSTONE, 1947).

Dentre as diversas abordagens para realização da rotação, a *varimax* é comumente utilizada e tem apresentado resultados satisfatórios (DE ALMEIDA *et al.*, 2020). O método busca maximizar a Eq. (14), sendo que a relação entre a i-ésima comunalidade e o *loading* do fator sendo rotacionado é dada por $\tilde{l}_{ij}^{\circ} = l_{ij}^{\circ} / \sqrt{h_i^2}$.

$$varimax = \frac{1}{p}\sum_{j=1}^{m}\left[\sum_{i=1}^{p} \tilde{l}_{ij}^{\circ\,4} - \frac{\left(\sum_{i=1}^{p} \tilde{l}_{ij}^{\circ\,2}\right)^2}{p}\right] \tag{14}$$

Maiores detalhes sobre este tipo de análise podem ser encontrados em (JOHNSON; WICHERN, 2007).

3. MATERIAIS E MÉTODOS

A presente seção busca elucidar quais foram os principais materiais e métodos utilizados durante o desenvolvimento desta pesquisa. Dessa forma, a fim de facilitar a compreensão do leitor, ela se encontra dividida em 3 subseções, sendo elas: caracterização da pesquisa, coleta de dados e métodos de análise.

3.1. Caracterização da pesquisa

O presente trabalho possui como objetivo a criação de um modelo de previsão da ocorrência de feições oleosas durante o processamento primário de petróleo. Trata-se de um problema de classificação binário, ou seja, apenas duas classes de resposta são possíveis, ocorrência ou não de feição para determinado cenário considerado. É importante destacar que nesses tipos de problemas a probabilidade associada a cada uma das classes é armazenada e pode, inclusive, ser utilizada para realização de uma análise fundamentalista das variáveis preditoras consideradas no processo.

Dessa forma, esta pesquisa caracteriza-se pela sua natureza aplicada devido ao seu interesse prático, isto é, os resultados serão aplicados ou utilizados imediatamente na solução de problemas reais. Em relação aos objetivos, ela pode ser classificada como normativa. Neste caso, primariamente, tem-se interesse no desenvolvimento de políticas, estratégias e ações para aperfeiçoar os resultados disponíveis na literatura existente, para encontrar uma solução ótima para novas definições de problemas ou para comparar várias estratégias relativas a um problema específico (BERTRAND; FRANSOO, 2002). Além disso, possui uma abordagem quantitativa, o que significa traduzir, em números, opiniões e informações a fim de classificá-las e analisá-las, através de técnicas estatísticas.

3.2. Coleta de dados

A técnica para a coleta de dados será principalmente a observação estruturada ou sistemática. Esta técnica realiza-se em condições controladas, para responder a propósitos preestabelecidos. Para gerar uma informação sobre a probabilidade de ocorrência de feição, por exemplo, uma organização das informações para que o estudo seja conduzido corretamente é requerida.

Os dados utilizados são advindos de uma plataforma *offshore* que realiza o processamento primário de petróleo e são compostos por 6 variáveis preditoras meteo-oceanográficas (DV, IV, DC, IC, DO e PP), além do TOG Espectrofotométrico. A base

de dados originais era inicialmente composta por 595 dados, dentre eles 150 apresentaram feição oleosa e 445 não apresentaram.

Convém destacar a importância de se trabalhar com uma base balanceada, uma vez que um problema muito comum é a utilização de uma base desbalanceada para treinamento do algoritmo de classificação aplicado para solucionar o problema (DEKAMIN *et al.*, 2021; GHADERYAN; ABBASI; SEDAAGHI, 2014; YU, HUALONG; NI; ZHAO, 2013). Uma base pode ser classificada como desbalanceada sempre que o número de observações pertencentes a uma determinada classe exceder o número de observações associadas a(s) outra(s) classe(s) (KOZIARSKI, 2020; LIN, WEI CHAO *et al.*, 2017).

De acordo com Koziarski (2020), muitos algoritmos comumente utilizados são suscetíveis a essa presença de dados desbalanceados. Esses algoritmos tendem a apresentar um viés em direção à classe predominante, perdendo, assim, sua capacidade de discriminação da classe minoritária. Surge, nesse contexto, uma grande dificuldade em se treinar um modelo adequado, uma vez que os modelos são dominados pela classe mais presente no conjunto de dados (LIN, WEI CHAO *et al.*, 2017; LIU, XU YING; WU; ZHOU, 2009).

Para facilitar a visualização de quão problemática pode ser essa situação, apresenta-se a seguir um exemplo encontrado em (LIN, WEI CHAO *et al.*, 2017). Toma-se um conjunto de dados em que 99% das observações pertencem à classe 0 e apenas 1% dos dados pertencem à classe 1. Dessa forma haverá maior tendência em classificar todos os dados como pertencentes à classe 0, já que com isso a acurácia do modelo será igual a 99%.

Dessa forma, a fim de trabalhar com uma base de dados balanceada, utilizou-se a técnica chamada de *undersampling* utilizando linguagem Python para tal. A Figura 17 mostra essa redução aplicada na base de dados. Sendo assim, ao todo foram considerados 300 observações coletados no período de 18 de abril de 2018 a 30 de julho de 2020, conforme pode ser observado no Anexo A deste trabalho. Convém destacar que esses dados foram coletados de acordo com o horário da foto obtida via satélite e por meio dessa foto que foi possível constatar a ocorrência ou não de feição.

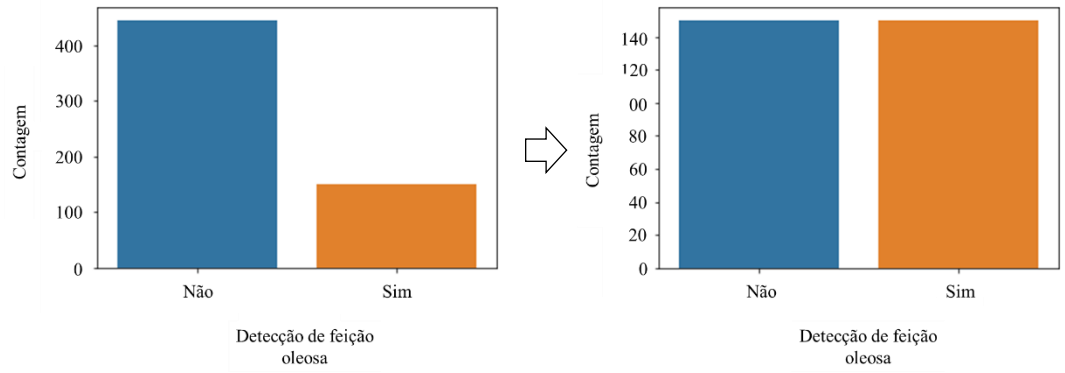


Figura 17. Síntese do processo de balanceamento dos dados

A principal forma de coleta de dados é via instrumentação nas plataformas de produção *offshore*, no caso do presente estudo. O Quadro 4 mostra quais os principais tipos de variáveis coletadas, a instrumentação utilizada e as variáveis consideradas para o presente estudo.

Quadro 4. Considerações sobre variáveis meteo-oceanográficas e instrumentos de medição.

Tipo da variável	Instrumentação	Variáveis mensuradas
Variáveis de vento	Principal instrumento utilizado é o anemômetro utilizado para mensurar a velocidade e a direção do vento.	Intensidade do vento Direção do vento
Variáveis de onda	Utilizam-se radares de ondas capaz de gerar um espectro de ondas da superfície do mar e então obter os valores das variáveis de interesse.	Período de pico Direção da onda
Variáveis de corrente	Principais instrumentos utilizados são sensores submersos, cuja localização mais comum é cerca de 15 a 20 metros de profundidade, podendo chegar a 40 metros de profundidade em alguns casos.	Intensidade da corrente Direção da corrente

Fonte: O autor

Para cada uma das variáveis mencionadas, obtém-se um valor a cada hora, sendo que o valor computado é o valor médio dos últimos 10 minutos de medição antes de completar a hora. Em relação ao vento, o anemômetro fica localizado em uma altura superior a 10 metros de altitude, conforme comumente realizado em terra. Entretanto, o valor é

convertido para velocidade a 10 metros de altitude por meio de correções matemáticas conforme apresentado no Guia para Instrumentos e Métodos de Observação (WMO, 2018).

Além disso, é importante destacar que a variável período de pico é obtida por meio de uma distribuição de valores associados ao comprimento das ondas intimamente ligados ao período da onda, ou seja, ao intervalo de tempo em que determinada onda passa. Assim, o período de pico é representado pelo período de pico que apresentar maior energia no mar em determinado instante, determinando assim o estado de mar. Períodos de pico mais curtos geram uma mistura muito maior, enquanto que períodos maiores promovem menores misturas.

Outra resposta avaliada neste estudo foi a extensão da feição oleosa, medida em milhas náuticas para os 150 casos em que a característica ocorreu e foi detectada. É importante realçar que a detecção dessas feições é efetuada por satélite e as medições das variáveis meteo-oceanográficas pelo centro de previsão meteorológica e estudos climáticos (centro de previsão de tempo e estudos climáticos - CPTEC). Além disso, o valor de TOG obtido corresponde ao TOG Espectrofotométrico, uma vez que este valor pode ser obtido em diferentes horários do dia para permitir uma reação proativa dos gestores da plataforma.

Nas Figuras de 18 a 23 são apresentadas algumas cartas de controle do tipo *Xbarra-R* para as variáveis meteo-oceanográficas. Para tal foram selecionadas observações para essas variáveis durante o período de 02 de janeiro de 2018 até 30 de agosto de 2020. Como algumas medições são realizadas em um período de 24 horas, utilizou-se a variável ‘dia’ como forma de definir o tamanho dos subgrupos.

Convém destacar que esses gráficos foram elaborados com o intuito de avaliar se existe variação entre os dias e durante o mesmo dia para as variáveis meteo-oceanográficas consideradas. Conforme esperado, já que são variáveis relacionadas com às condições naturais, pôde-se constatar que tanto os valores médios, apesentados na carta das médias amostrais, assim como os valores das amplitudes, mostrados na carta da amplitude apresentaram grandes variações. Dessa forma, podem-se observar valores bem distintos de intensidade do vento, por exemplo, durante a manhã e durante o período da tarde. Por essa razão, existe uma grande importância em monitorar essas variáveis constantemente e ser capaz de reagir com rapidez em momentos de elevados riscos.

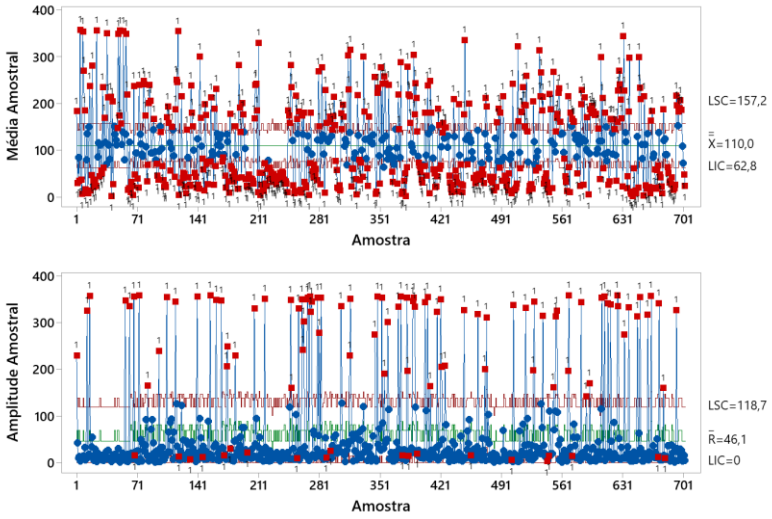


Figura 18. Carta Xbarra-R da variável Direção do Vento

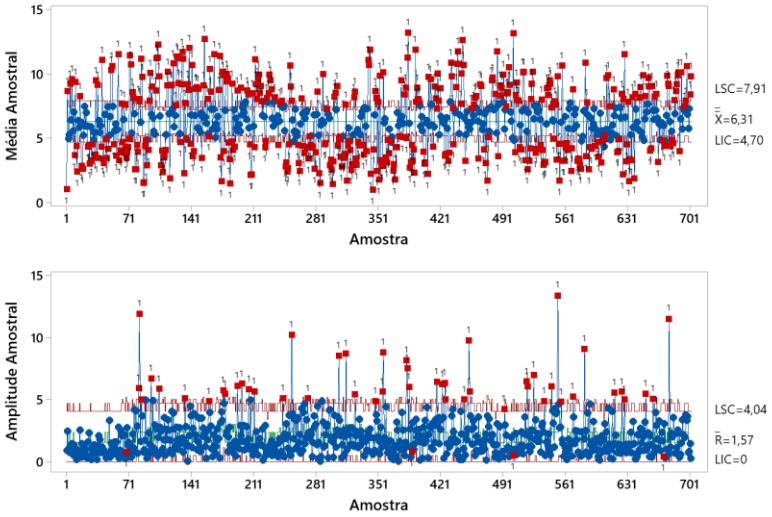


Figura 19. Carta Xbarra-R da variável Intensidade do Vento

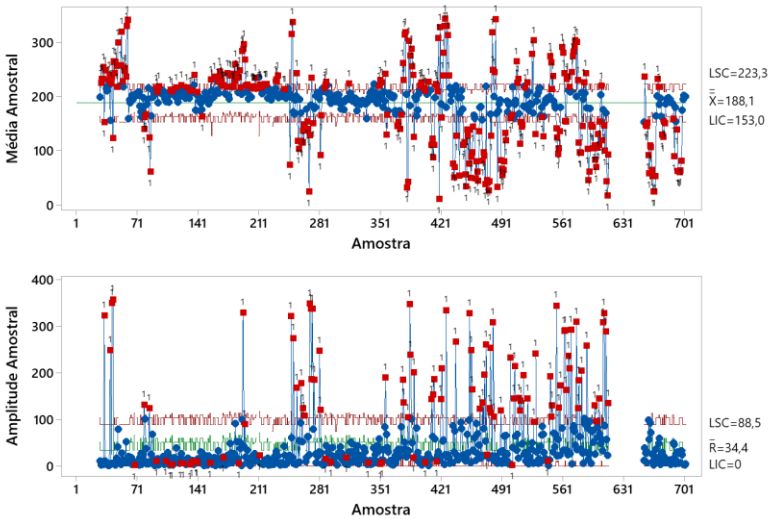


Figura 20. Carta Xbarra-R da variável Direção da Corrente

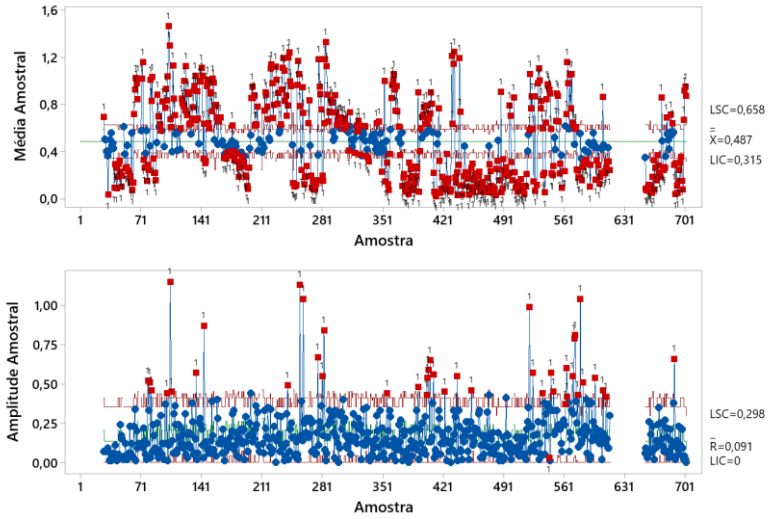


Figura 21. Carta Xbarra-R da variável Intensidade da Corrente

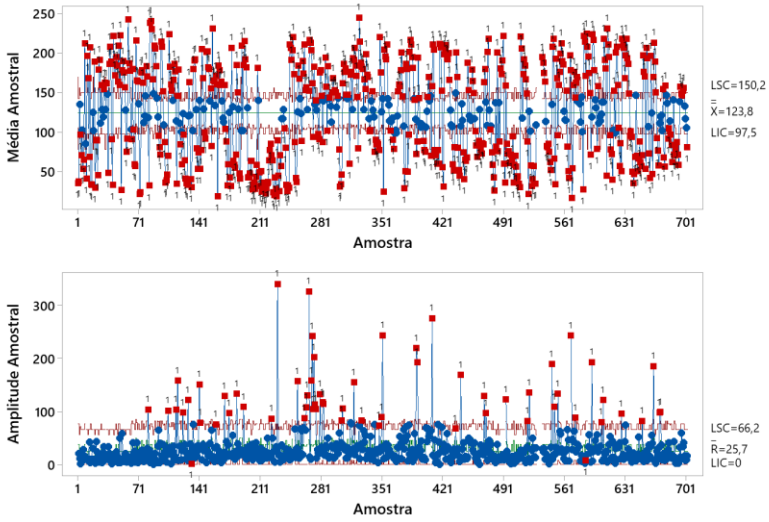


Figura 22. Carta Xbarra-R da variável Direção da onda

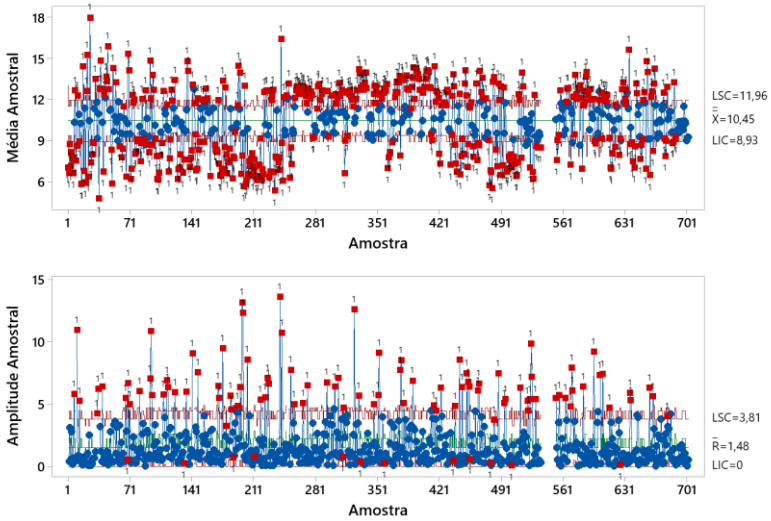


Figura 23. Carta Xbarra-R da variável Pico Primário

3.3. Métodos de análise

Ao todo são propostas 4 análises principais nesta pesquisa e cada uma delas será descrita de maneira sucinta a seguir.

A. Análise de risco associado aos cenários de feições

Esta primeira abordagem tem como objetivo avaliar a existência de cenários que estão mais propensos ao aparecimento de feições oleosas. A proposta é utilizar um arranjo fatorial completo, sendo o arranjo externo (fatorial completo 2^5) associado às variáveis meteo-oceanográficas com 32 experimentos e o interno (fatorial completo 2^1) associado às variáveis de carga poluidora, resultando em 2 experimentos. É importante destacar que os valores das variáveis superiores à mediana de cada uma das variáveis serão codificados como $+1$ e aqueles inferiores o serão como -1.

Assim, obtiveram-se 128 cenários formados pela combinação entre o arranjo interno (composto pela variável controlável) e arranjo externo (composto pelas variáveis incontroláveis). Para cada um deles, será contabilizada a quantidade de eventos com e sem presença de feição oleosa. Consequentemente a probabilidade de ocorrência de feição oleosa, P_f, pôde ser determinada, para cada cenário, por meio da Eq. (15), onde NF representa o número de ocorrência de feição e NS número de casos sem feição para um determinado cenário.

$$P_f = \frac{NF}{NF + NS} \tag{15}$$

Contudo, é importante mencionar que seria inconveniente associar uma probabilidade para ocorrência de feição de 0% para um cenário que aconteceu apenas uma vez e não houve detecção de feição. Assim, calcularam-se os valores de probabilidade de ocorrência de um cenário, P_c, por meio da quantidade de vezes em que esse cenário ocorreu em relação à quantidade total de 300 casos. Finalmente o risco para cada um dos cenários foi calculado por meio da multiplicação dessas duas probabilidades e foi convertido para uma escala de 0 a 1.

É importante destacar que para certas datas, havia valores faltantes para determinadas variáveis meteo-oceanográficas. Dessa forma, utilizou-se a estratégia de preenchimento desses valores faltantes pelos valores médios de suas respectivas colunas.

B. Desempenho dos classificadores

Neste estágio, o objetivo é realizar uma comparação entre possíveis classificadores a serem utilizados para prever a probabilidade de ocorrência e detecção de uma feição oleosa. Assim, o primeiro passo é definir quais classificadores serão utilizados, bem como os parâmetros associados a cada um deles.

Para uma melhor comparação, os conjuntos de treinamento e teste são divididos em 50 formas distintas, sempre respeitando a proporção dos dados de cada classe nos conjuntos de treinamento e teste. A seguir, é possível modelar e prever considerando todos os classificadores definidos na etapa anterior e todos os conjuntos de treinamento e teste obtidos. Métricas de desempenho, como acurácia, especificidade e sensibilidade, calculadas conforme mostrado na Eq. (1), Eq. (2) e Eq. (3), respectivamente, e seus valores médios são extraídos para fins de comparação.

Além dos classificadores individuais, é importante destacar que foram avaliados dois *ensembles*, ou seja, combinações dos resultados dos classificadores fornecendo pesos iguais a cada um deles. O primeiro ensemble foi elaborado por meio da ponderação da probabilidade de cada método individual, gerando assim a probabilidade final da observação para o *ensemble* e por meio desse valor, define-se a classe à qual o registro irá pertencer.

A segunda forma estruturar o *ensemble* de classificadores foi por meio da atribuição da classificação mais frequente, segundo os classificadores individuais, como sendo a classificação final para cada uma das observações do conjunto de dados. Essa última abordagem apresenta como desvantagem o fato de não se encontrar uma probabilidade associada ao *ensemble*, diferentemente do caso anterior. Caso ela fosse a melhor abordagem, haveria problemas em se utilizar a metodologia proposta no Estágio C.

Convém destacar que a mesma metodologia de avaliação de desempenho utilizada para os classificadores individuais foi utilizada para os *ensembles*. Assim, para cada conjunto de treinamento e teste armazenaram-se as métricas de acurácia, especificidade e sensibilidade e ao final a média foi armazenada.

C. Modelo para previsão da ocorrência e detecção de feições oleosas

O método selecionado na análise anterior é aplicado para modelar os dados de treinamento e fazer previsões para o conjunto de teste, e as probabilidades de ocorrência e detecção de feição oleosa devem ser armazenadas. É importante avaliar se os dados de

treinamento são adequados para realização da análise fatorial. Se os dados seguirem uma distribuição multivariada normal, pode-se usar o teste de esfericidade de Bartlett, que testa a hipótese de que as variáveis não são correlacionadas, por meio da comparação entre a matriz de correlações com a matriz identidade (FÁVERO, 2015).

Uma vez que os dados não sigam uma distribuição normal multivariada, utiliza-se a medida de adequação amostral Kaiser-Meyer-Olkin (KMO), responsável por avaliar se os dados são adequados para a realização da análise fatorial. Valores entre 0,5 e 1,0 para este índice indicam que a análise fatorial é aceitável. A análise fatorial é recomendada se pelo menos um dos testes anteriores for favorável. A Eq. (16) representa a fórmula utilizada para o cálculo do índice KMO, onde r_{ij} e q_{ij} representam, respectivamente, as matrizes de correlação de amostra **R** e anti-imagem **Q**. O valor para este indicador é superior a 1, sendo que um índice desejável é aquele superior a 0,5 (DE ALMEIDA *et al.*, 2020).

$$KMO = \frac{\sum_{i \neq j} r_{ij}^2}{\sum_{i \neq j} r_{ij}^2 + \sum_{i \neq j} q_{ij}^2} \tag{16}$$

Considerando um cenário onde a análise fatorial é recomendada, os escores dos fatores rotacionados são armazenados e considerados como variáveis preditoras. O conjunto de treinamento está agrupado em 2 grupos: o primeiro com apenas observações em que a mancha de óleo ocorreu e foi detectada e o segundo com observações sem ocorrência de mancha de óleo. Os dados de entrada são convertidos em uma matriz de superfície de resposta do tipo CCD para avaliar os efeitos principais, interações e efeitos quadráticos relacionados aos fatores que representam as variáveis originais.

Em seguida, a modelagem das probabilidades é realizada selecionando os termos significativos do modelo. Um algoritmo de otimização é utilizado neste caso para verificar os efeitos de cada variável, uma vez que as interações e os termos quadráticos podem ser significativos, o que significa que apenas a análise dos efeitos principais pode levar a conclusões errôneas. Gráficos de contorno e superfície de resposta também são utilizados nesse estágio para facilitar a compreensão dos efeitos dessas variáveis.

D. Modelagem da extensão da feição oleosa

Considerando apenas os dados em que foram detectadas feições oleosas e medida sua extensão, avalia-se novamente a necessidade de realização de uma análise fatorial, mas,

neste caso, verificou-se que o teste KMO não apresentou valores superiores a 0,5 para a maioria parte das variáveis, optando, portanto, em não realizar a análise fatorial nesse contexto. Em seguida, os preditores originais são convertidos em um arranjo do tipo fatorial customizado e a extensão da feição oleosa é analisada como resposta desse arranjo.

Nesta análise, optou-se por usar um projeto fatorial completo porque os efeitos quadráticos e as interações não melhoraram a qualidade do modelo, além de reduzirem o valor da métrica R^2_{adj}. Os efeitos das variáveis meteo-oceanográficas e do TOG na extensão da feição nessa etapa são avaliados por meio dos gráficos de efeitos principais, dos gráficos de contorno e também da superfície de resposta.

As etapas de todos os 4 estágios da metodologia aqui apresentados estão sintetizadas na Figura 24 e na Figura 25 para uma melhor compreensão.

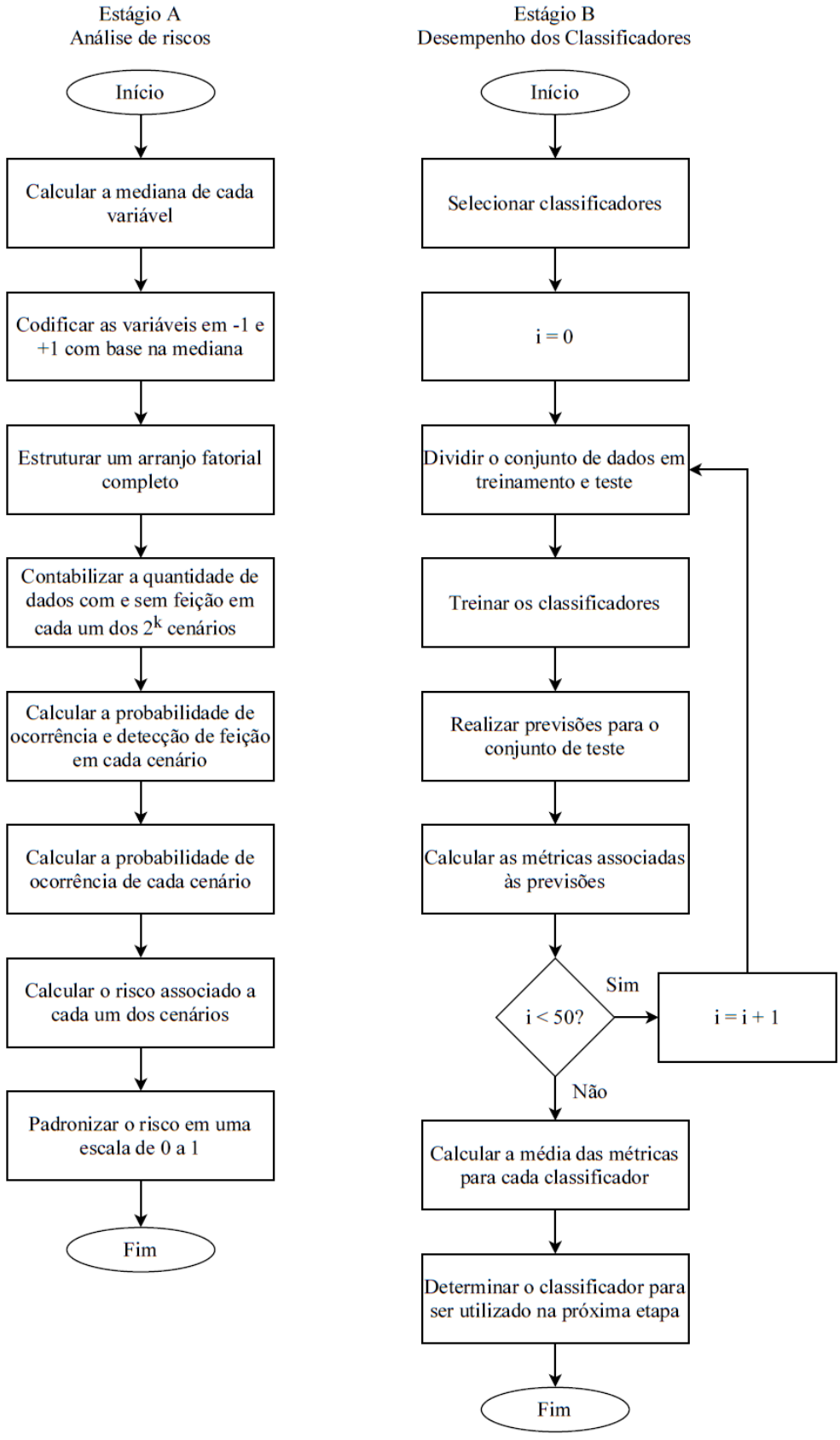


Figura 24. Fluxograma dos estágios A e B da metodologia utilizada neste estudo

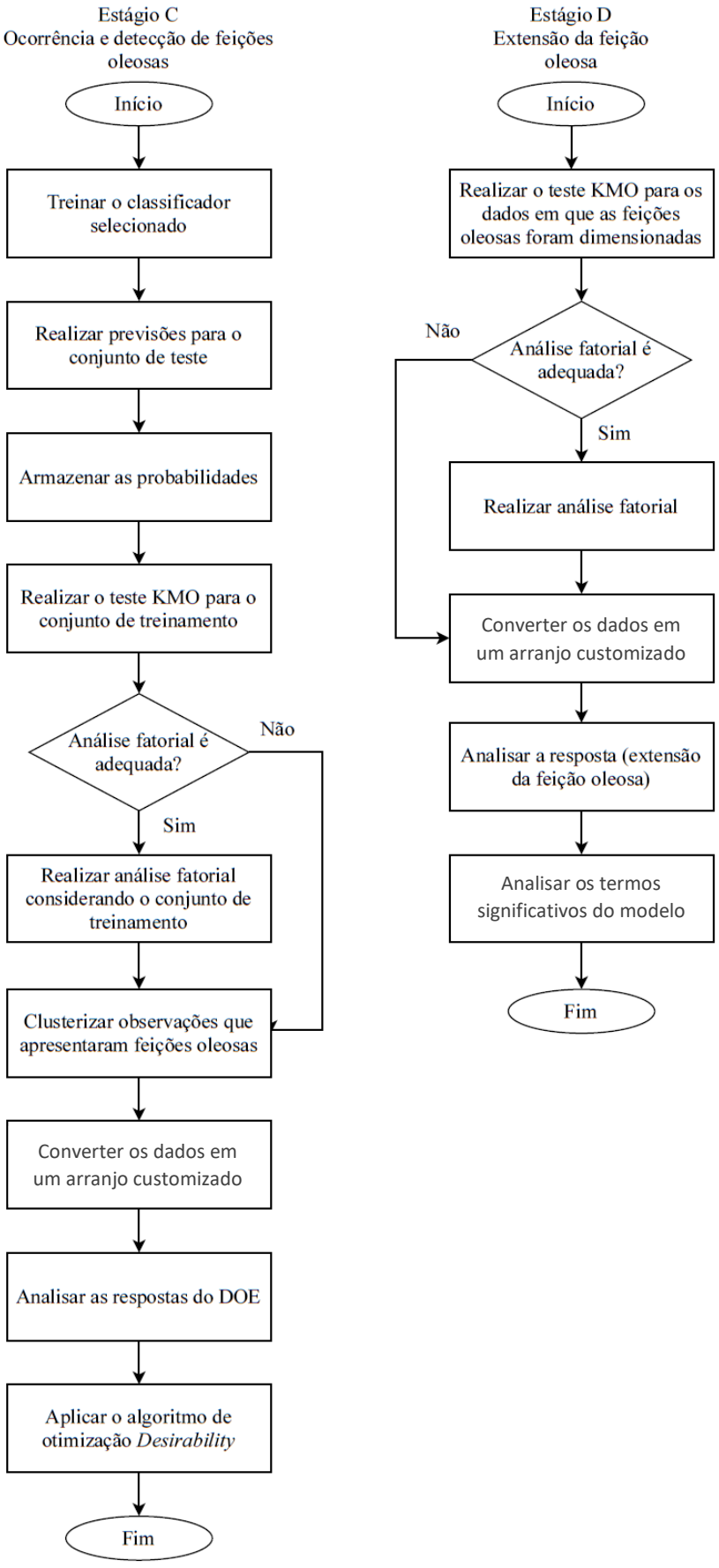


Figura 25. Fluxograma dos estágios C e D da metodologia utilizada neste estudo

4. RESULTADOS E DISCUSSÕES

Nesta seção, serão apresentados e discutidos os principais resultados obtidos em cada uma das análises propostas para o problema de pesquisa do presente trabalho. Além de algumas observações feitas a partir de algumas análises preliminares.

4.1. Análises preliminares

A fim de possibilitar maior compreensão a respeito da complexidade do problema de classificação proposto no presente trabalho, elaborou-se a Figura 26. Nela é possível verificar as observações em verde indicando casos em que não houve ocorrência de feição oleosa e em laranja fazendo referência aos casos em que a presença de feição oleosa pode ser constatada.

Figura 26. Gráfico das observações separadas por classes em função das variáveis de entrada

A Tabela 2 traz maiores detalhes a respeito dessas variáveis, como: valores máximos e mínimos, mediana, bem como suas respectivas unidades de medida.

Tabela 2. Estatísticas descritivas relevantes para análises futuras

Tipo	Variáveis	Mínimo	Máximo	Mediana
Incontroláveis	Direção do vento (°)	0,3	358,2	93,2
	Intensidade do vento (m/s)	0,57	13,8	4,84
	Direção corrente (°)	4,46	357,89	201,77
	Intensidade corrente (m/s)	0,01	1,45	0,48
	Direção da onda (°)	10,34	342,88	134,12
	Período pico primário (s)	4,94	18,04	10,60
Controlável	TOG (mg/L)	3,5	28,31	11,17

As Figura 27 Figura 28 mostram os *boxplots* para os dados considerados nesse estudo, que também podem ser observados na Tabela A 1, que se encontra no Anexo A desse documento. Por meio dessas figuras é possível observar a variabilidade das variáveis preditoras para dois níveis, com e sem feição.

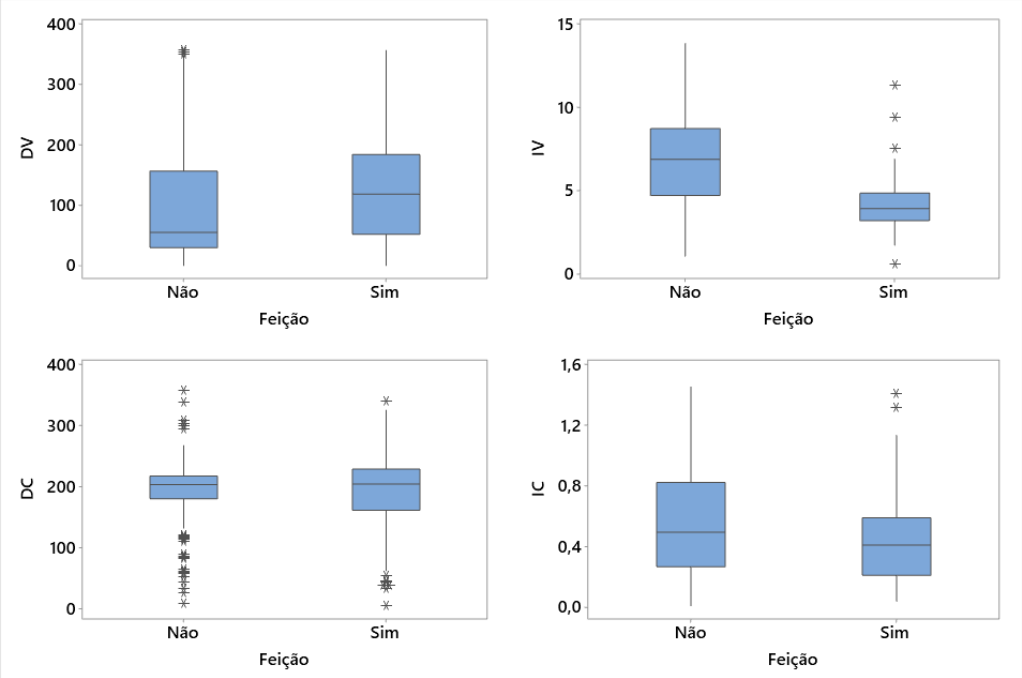


Figura 27. *Boxplots* das variáveis meteo-oceanográficas e TOG considerando os dados analisados no estudo (Parte I)

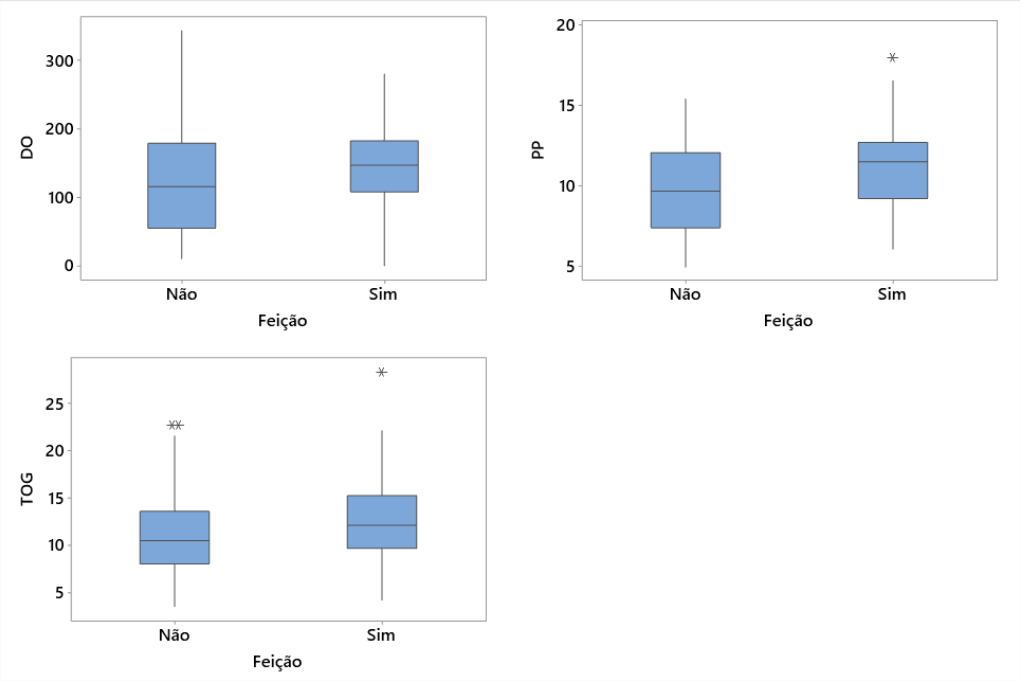


Figura 28. *Boxplots* das variáveis meteo-oceanográficas e TOG considerando os dados analisados no estudo (Parte II)

4.2. Análise de riscos

No estágio A foi realizada a análise dos riscos associados a ocorrência de feição de acordo com os cenários existentes. Conforme explicado na seção 3 deste trabalho, os valores das variáveis controláveis e não controláveis foram classificados em acima ou abaixo da mediana, dando origem à Tabela 3.

Tabela 3. Quantidade de casos com e sem ocorrência feição, probabilidade de ocorrência e risco associados aos 128 possíveis cenários (Parte I)

Variáveis meteo-oceanográficas e TOG							**Casos**		**Probabilidades e risco**		
DV	IV	DC	IC	DO	PPP	TOG	Sem feição	Com feição	P_f	P_C	Risco 0 – 1
-1	-1	-1	-1	-1	-1	-1	0	2	1,00	0,01	0,38
1	-1	-1	-1	-1	-1	-1	1	0	0,00	0,00	0,00
-1	1	-1	-1	-1	-1	-1	2	0	0,00	0,00	0,00
1	1	-1	-1	-1	-1	-1	1	0	0,00	0,00	0,00
-1	-1	1	-1	-1	-1	-1	1	0	0,00	0,00	0,00
1	-1	1	-1	-1	-1	-1	2	6	0,75	0,02	0,63
-1	1	1	-1	-1	-1	-1	1	1	0,50	0,01	0,09
1	1	1	-1	-1	-1	-1	0	3	1,00	0,01	0,50
-1	-1	-1	1	-1	-1	-1	0	1	1,00	0,01	0,25
1	-1	-1	1	-1	-1	-1	0	2	1,00	0,01	0,38
-1	1	-1	1	-1	-1	-1	10	1	0,09	0,00	0,01
1	1	-1	1	-1	-1	-1	0	1	1,00	0,01	0,25
-1	-1	1	1	-1	-1	-1	3	0	0,00	0,00	0,00

Continuação da Tabela 3

Variáveis meteo-oceanográficas e TOG							Casos		Probabilidades e risco		
DV	IV	DC	IC	DO	PPP	TOG	Sem feição	Com feição	P_f	P_C	Risco 0 – 1
1	-1	1	1	-1	-1	-1	0	1	1,00	0,01	0,25
-1	1	1	1	-1	-1	-1	14	2	0,13	0,01	0,03
1	1	1	1	-1	-1	-1	1	1	0,50	0,01	0,09
-1	-1	-1	-1	1	-1	-1	1	1	0,50	0,01	0,09
1	-1	-1	-1	1	-1	-1	5	0	0,00	0,00	0,00
-1	1	-1	-1	1	-1	-1	0	0	-	-	-
1	1	-1	-1	1	-1	-1	2	0	0,00	0,00	0,00
-1	-1	1	-1	1	-1	-1	0	1	1,00	0,01	0,25
1	-1	1	-1	1	-1	-1	1	4	0,80	0,02	0,48
-1	1	1	-1	1	-1	-1	0	0	-	-	-
1	1	1	-1	1	-1	-1	1	0	0,00	0,00	0,00
-1	-1	-1	1	1	-1	-1	2	0	0,00	0,00	0,00
1	-1	-1	1	1	-1	-1	0	1	1,00	0,01	0,25
-1	1	-1	1	1	-1	-1	0	0	-	-	-
1	1	-1	1	1	-1	-1	1	0	0,00	0,00	0,00
-1	-1	1	1	1	-1	-1	0	1	1,00	0,01	0,25
1	-1	1	1	1	-1	-1	2	2	0,50	0,01	0,16
-1	1	1	1	1	-1	-1	0	0	-	-	-
1	1	1	1	1	-1	-1	1	0	0,00	0,00	0,00
-1	-1	-1	-1	-1	1	-1	1	0	0,00	0,00	0,00
1	-1	-1	-1	-1	1	-1	0	1	1,00	0,01	0,25
-1	1	-1	-1	-1	1	-1	1	0	0,00	0,00	0,00
1	1	-1	-1	-1	1	-1	0	0	-	-	-
-1	-1	1	-1	-1	1	-1	1	0	0,00	0,00	0,00
1	-1	1	-1	-1	1	-1	0	0	-	-	-
-1	1	1	-1	-1	1	-1	0	2	1,00	0,01	0,38
1	1	1	-1	-1	1	-1	0	0	-	-	-
-1	-1	-1	1	-1	1	-1	0	0	-	-	-
1	-1	-1	1	-1	1	-1	0	1	1,00	0,01	0,25
-1	1	-1	1	-1	1	-1	3	1	0,25	0,00	0,04
1	1	-1	1	-1	1	-1	0	0	-	-	-
-1	-1	1	1	-1	1	-1	1	0	0,00	0,00	0,00
1	-1	1	1	-1	1	-1	0	0	-	-	-
-1	1	1	1	-1	1	-1	7	1	0,13	0,00	0,02
1	1	1	1	-1	1	-1	1	0	0,00	0,00	0,00
-1	-1	-1	-1	1	1	-1	0	1	1,00	0,01	0,25
1	-1	-1	-1	1	1	-1	2	0	0,00	0,00	0,00
-1	1	-1	-1	1	1	-1	3	1	0,25	0,00	0,04
1	1	-1	-1	1	1	-1	5	1	0,17	0,00	0,02
-1	-1	1	-1	1	1	-1	0	0	-	-	-
1	-1	1	-1	1	1	-1	0	4	1,00	0,02	0,63
-1	1	1	-1	1	1	-1	2	1	0,33	0,00	0,06
1	1	1	-1	1	1	-1	6	0	0,00	0,00	0,00

Continuação da Tabela 3

Variáveis meteo-oceanográficas e TOG							Casos		Probabilidades e risco		
DV	IV	DC	IC	DO	PPP	TOG	Sem feição	Com feição	P_f	P_C	Risco 0 – 1
-1	-1	-1	1	1	1	-1	1	1	0,50	0,01	0,09
1	-1	-1	1	1	1	-1	0	7	1,00	0,03	1,00
-1	1	-1	1	1	1	-1	0	1	1,00	0,01	0,25
1	1	-1	1	1	1	-1	0	2	1,00	0,01	0,38
-1	-1	1	1	1	1	-1	2	1	0,33	0,00	0,06
1	-1	1	1	1	1	-1	0	5	1,00	0,02	0,75
-1	1	1	1	1	1	-1	0	2	1,00	0,01	0,38
1	1	1	1	1	1	-1	2	0	0,00	0,00	0,00

Tabela 4. Quantidade de casos com e sem ocorrência feição, probabilidade de ocorrência e risco associados aos 128 possíveis cenários (Parte II)

Variáveis meteo-oceanográficas e TOG							Casos		Probabilidades e risco		
DV	IV	DC	IC	DO	PPP	TOG	Sem feição	Com feição	P_f	P_C	Risco 0 – 1
-1	-1	-1	-1	-1	-1	1	0	3	1,00	0,01	0,50
1	-1	-1	-1	-1	-1	1	1	3	0,75	0,01	0,35
-1	1	-1	-1	-1	-1	1	0	0	-	-	-
1	1	-1	-1	-1	-1	1	0	1	1,00	0,01	0,25
-1	-1	1	-1	-1	-1	1	1	1	0,50	0,01	0,09
1	-1	1	-1	-1	-1	1	0	2	1,00	0,01	0,38
-1	1	1	-1	-1	-1	1	6	3	0,33	0,01	0,14
1	1	1	-1	-1	-1	1	1	0	0,00	0,00	0,00
-1	-1	-1	1	-1	-1	1	0	2	1,00	0,01	0,38
1	-1	-1	1	-1	-1	1	0	0	-	-	-
-1	1	-1	1	-1	-1	1	4	2	0,33	0,01	0,10
1	1	-1	1	-1	-1	1	1	0	0,00	0,00	0,00
-1	-1	1	1	-1	-1	1	0	0	-	-	-
1	-1	1	1	-1	-1	1	0	2	1,00	0,01	0,38
-1	1	1	1	-1	-1	1	8	3	0,27	0,01	0,11
1	1	1	1	-1	-1	1	0	1	1,00	0,01	0,25
-1	-1	-1	-1	1	-1	1	0	0	-	-	-
1	-1	-1	-1	1	-1	1	0	7	1,00	0,03	1,00
-1	1	-1	-1	1	-1	1	0	0	-	-	-
1	1	-1	-1	1	-1	1	1	0	0,00	0,00	0,00
-1	-1	1	-1	1	-1	1	0	0	-	-	-
1	-1	1	-1	1	-1	1	1	2	0,67	0,01	0,22
-1	1	1	-1	1	-1	1	0	0	-	-	-
1	1	1	-1	1	-1	1	1	1	0,50	0,01	0,09
-1	-1	-1	1	1	-1	1	0	1	1,00	0,01	0,25
1	-1	-1	1	1	-1	1	2	3	0,60	0,01	0,27
-1	1	-1	1	1	-1	1	0	0	-	-	-

Continuação da Tabela 4

Variáveis meteo-oceanográficas e TOG							Casos		Probabilidades e risco		
DV	IV	DC	IC	DO	PPP	TOG	Sem feição	Com feição	P_f	P_C	Risco 0 – 1
1	1	-1	1	1	-1	1	2	0	0,00	0,00	0,00
-1	-1	1	1	1	-1	1	1	0	0,00	0,00	0,00
1	-1	1	1	1	-1	1	1	0	0,00	0,00	0,00
-1	1	1	1	1	-1	1	0	0	-	-	-
1	1	1	1	1	-1	1	0	0	-	-	-
-1	-1	-1	-1	-1	1	1	0	1	1,00	0,01	0,25
1	-1	-1	-1	-1	1	1	2	6	0,75	0,02	0,63
-1	1	-1	-1	-1	1	1	3	2	0,40	0,01	0,12
1	1	-1	-1	-1	1	1	0	0	-	-	-
-1	-1	1	-1	-1	1	1	1	0	0,00	0,00	0,00
1	-1	1	-1	-1	1	1	0	3	1,00	0,01	0,50
-1	1	1	-1	-1	1	1	0	0	-	-	-
1	1	1	-1	-1	1	1	1	0	0,00	0,00	0,00
-1	-1	-1	1	-1	1	1	0	0	-	-	-
1	-1	-1	1	-1	1	1	0	1	1,00	0,01	0,25
-1	1	-1	1	-1	1	1	2	1	0,33	0,00	0,06
1	1	-1	1	-1	1	1	0	0	-	-	-
-1	-1	1	1	-1	1	1	0	0	-	-	-
1	-1	1	1	-1	1	1	0	1	1,00	0,01	0,25
-1	1	1	1	-1	1	1	1	1	0,50	0,01	0,09
1	1	1	1	-1	1	1	1	0	0,00	0,00	0,00
-1	-1	-1	-1	1	1	1	1	3	0,75	0,01	0,35
1	-1	-1	-1	1	1	1	4	5	0,56	0,02	0,39
-1	1	-1	-1	1	1	1	1	2	0,67	0,01	0,22
1	1	-1	-1	1	1	1	4	3	0,43	0,01	0,18
-1	-1	1	-1	1	1	1	0	5	1,00	0,02	0,75
1	-1	1	-1	1	1	1	1	1	0,50	0,01	0,09
-1	1	1	-1	1	1	1	1	0	0,00	0,00	0,00
1	1	1	-1	1	1	1	1	0	0,00	0,00	0,00
-1	-1	-1	1	1	1	1	0	4	1,00	0,02	0,63
1	-1	-1	1	1	1	1	0	2	1,00	0,01	0,38
-1	1	-1	1	1	1	1	2	0	0,00	0,00	0,00
1	1	-1	1	1	1	1	1	0	0,00	0,00	0,00
-1	-1	1	1	1	1	1	0	5	1,00	0,02	0,75
1	-1	1	1	1	1	1	0	3	1,00	0,01	0,50
-1	1	1	1	1	1	1	1	0	0,00	0,00	0,00
1	1	1	1	1	1	1	1	0	0,00	0,00	0,00

A partir das análises realizadas, é possível identificar que existe uma diferença entre o valor do risco associado a cada um dos cenários. Dessa forma, o fenômeno não é aleatório, uma vez que de fato existem casos em que há uma maior propensão para o aparecimento de feições oleosas. É importante ressaltar que existem cenários com grande

risco de ocorrência de feição mesmo para casos em que o TOG esteve em um nível mais baixo, revelando um indício de que as variáveis meteo-oceanográficas terão uma grande importância nesse problema. Nas próximas subseções, análises mais profundas acerca da importância de cada uma das variáveis serão apresentadas.

4.3. Desempenho dos classificadores

Neste trabalho, inicialmente 5 métodos comumente aplicados foram selecionados, a saber, RF, KNN, MLP, RLB e SVM. Executaram-se 50 conjuntos diferentes de treinamento e teste usando a função *kfold.split* disponível em linguagem Python, com 535 observações para treinamento e 60 para teste. Consequentemente, foram criados 50 modelos diferentes para cada um desses métodos, de forma que a acurácia de cada um fosse computada a cada etapa e a média de todas essas execuções fosse armazenada. A Tabela 5 mostra os métodos utilizados, a parametrização de cada um deles e as métricas obtidas (em média) para o conjunto de teste.

Tabela 5. Parâmetros e métricas associadas às técnicas de *machine learning* utilizadas

Técnica de *Machine learning*	Parâmetros	A_c	S_p	S_n
RF	Número de estimadores = 150 Número máximo de preditores = 3	0,77	0,73	0,82
KNN	k = 5	0,74	0,65	0,84
MLP	Número de camadas escondidas = 2 Número de neurônios = 4 Função de ativação = *Relu* Solver = LBFGS α = 0,05 Taxa de aprendizado = invscaling	0,71	0,68	0,76
RLB	*Link function* = logit	0,74	0,70	0,80
SVM	*Kernel* = RBF γ = 0.1 C = 1	0,75	0,65	0,85

Em termos de acurácia (A_c), o RF foi o método que apresentou melhor desempenho. Assim, ele foi selecionado para ser usado na subseção 4.2. Para um melhor entendimento da influência das variáveis na ocorrência e detecção de feições oleosas, foi armazenado o gráfico de importância dos preditores (*feature importance*), obtido pelo RF, conforme mostrado na Figura 29.

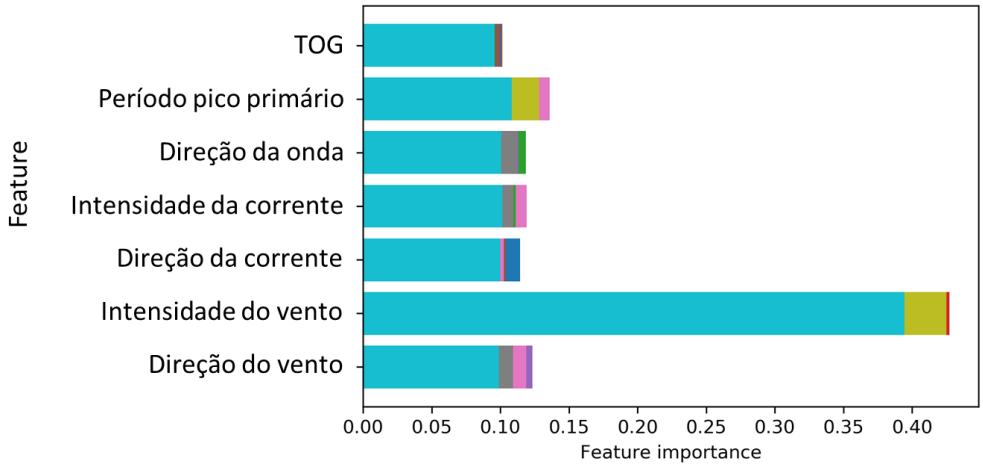


Figura 29. Gráfico de *feature importance* para as 50 execuções do RF

É importante destacar que para todos os casos, considerou-se um limiar de 50%, ou seja, registros cuja probabilidade associada estimada ou prevista pelo modelo apresentar valores superiores a esse limiar serão classificados como classe 1, do contrário serão classificados como classe 0.

Uma outra alternativa seria considerar uma combinação de classificadores (*ensemble*), demonstrada por alguns trabalhos na literatura como sendo de imensa importância para melhorar a qualidade das previsões (CHEN, HUAZHOU *et al.*, 2018; LAZRI; AMEUR, 2018; MANSOURI *et al.*, 2013). A partir disso, avaliou-se a combinação dos 3 melhores classificadores, ou seja, RF, RLB e SVM. As probabilidades associadas às previsões de cada um desses métodos para cada um dos 50 conjuntos de teste foram armazenadas. Em seguida, esses valores foram ponderados com o mesmo valor de peso ($1/3$, ou seja, 33,33%) e somados.

O valor final encontrado foi tido como o valor da probabilidade associada ao *ensemble* para cada uma das observações do conjunto de teste. Nessa etapa o limiar adotado também foi de 50%. As métricas de acurácia, especificidade e sensibilidade foram calculadas para os 50 conjuntos de teste e, então, os valores dessas métricas para o ensemble foram comparados, por meio do teste t pareado (*paired t*), com os valores encontrados para os métodos RF, RLB e SVM.

Ao final desta análise, concluiu-se que o método RF ainda apresentava melhores resultados em termos das métricas avaliadas. Os *p-values* encontrados em cada um dos testes estão apresentados na Tabela 6. Valores de *p-value* inferiores a 0,05 indicam que a

hipótese de igualdade entre as médias deve ser rejeitada, uma vez que todas as análises realizadas nessa pesquisa consideraram um nível de significância igual a 5%. Convém mencionar que os testes foram realizados considerando a hipótese de igualdade entre as médias e que esses podem ser vistos com maiores detalhes no Apêndice desse trabalho.

Tabela 6. Valores de *p-value* para cada um dos testes t pareado executados para o *ensemble* 1

	RF $\mathbf{Ac_{média} = 78,74\%}$	**RLB** $\mathbf{Ac_{média} = 74,73\%}$	**SVM** $\mathbf{Ac_{média} = 76,80\%}$
Ensemble $\mathbf{Ac_{média} = 77,73\%}$	0,345	0,001	0,204
	RF $\mathbf{Sp_{média} = 74,73\%}$	**RLB** $\mathbf{Sp_{média} = 71,92\%}$	**SVM** $\mathbf{Sp_{média} = 70,18\%}$
Ensemble $\mathbf{Sp_{média} = 70,91\%}$	0,001	0,404	0,400
	RF $\mathbf{Sn_{média} = 81,22\%}$	**RLB** $\mathbf{Sn_{média} = 76,76\%}$	**SVM** $\mathbf{Sn_{média} = 82,37\%}$
Ensemble $\mathbf{Sn_{média} = 83,16\%}$	0,053	0,000	0,474

Ao analisar a Tabela 6, é possível constatar que os valores de acurácia e sensibilidade média do ensemble não diferem dos valores médios para essas métricas associadas ao RF, conforme indicado pelos *p-values*. O valor da especificidade apresenta diferença e a média apresentada pelo RF é superior.

O *ensemble* apresentou desempenho superior à regressão logística binária, uma vez que há diferença entre a acurácia e sensibilidade médias e em ambos os casos as médias associadas ao ensemble foram superiores. Considerando a especificidade, ambos os métodos apresentaram médias estatisticamente iguais. Finalmente, ao comparar o desempenho do *ensemble* com o desempenho do *support vector machine*, todas as métricas apresentaram valores que não diferem estatisticamente, conforme indicado pelos *p-values* apresentados.

Uma segunda abordagem para a combinação de classificadores também foi elaborada no presente trabalho. Nesse momento considerou-se a previsão final do ensemble como sendo a classe mais frequentemente entre os 3 melhores classificadores selecionados anteriormente para cada observação de cada um dos 50 conjuntos de teste. Os resultados foram bem similares aos encontrados na primeira abordagem, entretanto algumas diferenças em alguns casos puderam ser observadas, o que pode ser constatado comparando-se os valores médios das métricas de desempenho consideradas para os

ensembles na primeira e segunda abordagens apresentadas na Tabela 6 e Tabela 7, respectivamente.

A Tabela 7 apresenta os valores de *p-value* para os testes t pareado executados na comparação das médias das métricas avaliadas considerando a segunda abordagem para o ensemble e os métodos individuais. Novamente considerou-se um nível de significância igual a 5% para a realização dos testes.

Tabela 7. Valores de *p-value* para cada um dos testes t pareado executados para o ensemble 2

	RF $\mathbf{Ac_{média} = 78,74\%}$	**RLB** $\mathbf{Ac_{média} = 74,73\%}$	**SVM** $\mathbf{Ac_{média} = 76,80\%}$
Ensemble $\mathbf{Ac_{média} = 77,13\%}$	0,048	0,002	0,651
	RF $\mathbf{Sp_{média} = 74,73\%}$	**RLB** $\mathbf{Sp_{média} = 71,92\%}$	**SVM** $\mathbf{Sp_{média} = 70,18\%}$
Ensemble $\mathbf{Sp_{média} = 70,02\%}$	0,000	0,112	0,833
	RF $\mathbf{Sn_{média} = 81,22\%}$	**RLB** $\mathbf{Sn_{média} = 76,76\%}$	**SVM** $\mathbf{Sn_{média} = 82,37\%}$
Ensemble $\mathbf{Sn_{média} = 82,74\%}$	0,082	0,000	0,736

Dessa forma, o método RF continua sendo o que apresenta melhor desempenho, mesmo em relação às duas abordagens de *ensemble*, para a análise dos dados utilizados nessa pesquisa e por essa razão foi selecionado para ser utilizado nas etapas posteriores desse trabalho. É importante destacar que as médias para as métricas A_c, S_p e S_n apresentadas durante a comparação dos métodos individuais com o *ensemble*, não são exatamente iguais àquelas apresentadas na Tabela 5, uma vez que se utilizou outro código ainda utilizando linguagem Python, dessa forma as divisões do conjunto de dados foram diferentes da primeira análise. Entretanto, é possível observar que o comportamento permanece, ou seja, o RF permanece como sendo o método de maior acurácia, seguido pelo SVM e pela RLB, respectivamente. O mesmo pode ser observado em relação às outras técnicas.

4.4. Modelo probabilístico para ocorrência e detecção de feição

Dividindo-se o conjunto novamente em treinamento e teste, seguindo a mesma proporção mencionada anteriormente, e aplicando o método *Random Forest*, pelas razões anteriormente explicadas, foi obtida a seguinte matriz de confusão para o algoritmo, conforme mostrada na Tabela 8. Os acertos estão apresentados na diagonal principal,

enquanto que as classificações errôneas estão na diagonal secundária. Num total de 30 dados o modelo acertou 27 casos, o que resulta em um valor de acurácia igual a 90,00%. Nos casos em que não havia feição, ou seja, 14 casos, o modelo previu corretamente 12, levando a um valor de especificidade de 85,71%. Por fim, em relação à sensibilidade, ou seja, o número de casos em que houve feição que o modelo previu corretamente, o modelo apresentou 93,75%.

Tabela 8. Matriz de confusão para o modelo RF

		PREVISÃO	
		Classe 0	*Classe 1*
REAL	*Classe 0*	12	2
	Classe 1	1	15

O gráfico de importância dos preditores para o modelo pode ser observado na Figura 30. Assim, é possível observar que a variável intensidade do vento possui maior importância dentre as preditoras (valor superior a 30%), enquanto que as demais variáveis apresentaram importância na faixa de 10%. A

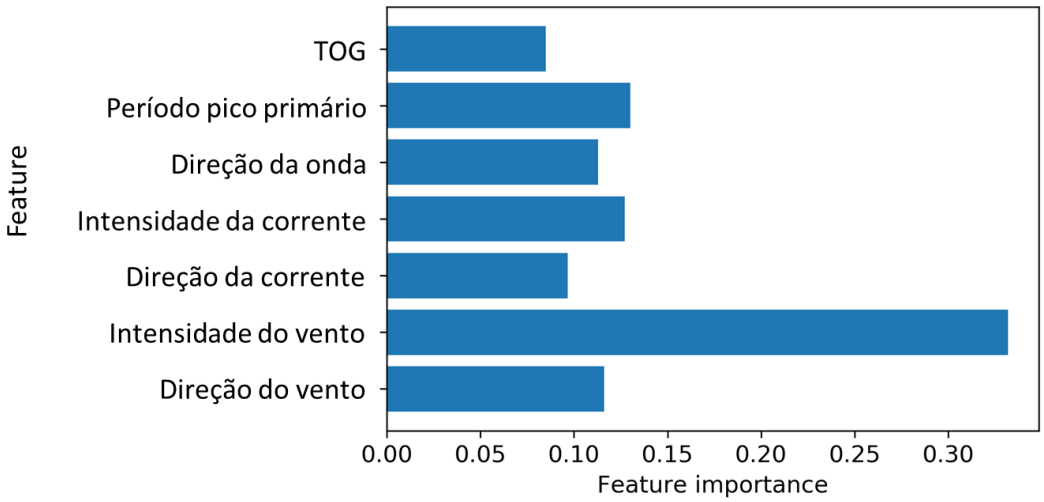


Figura 30. Gráfico de *feature importance* para o modelo RF

A Figura 31 mostra a curva ROC para o modelo, apresentando um valor de área sob a curva igual a 93%, que juntamente com as métricas avaliadas pela matriz de confusão, reforça a validade no modelo preditivo de ocorrência de feições oleosas por meio do algoritmo RF.

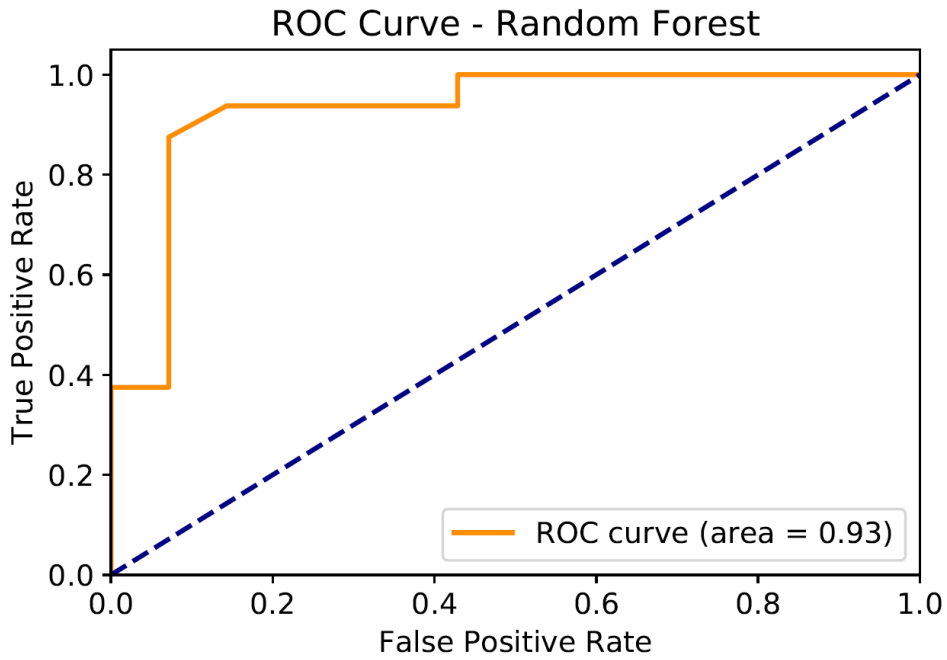


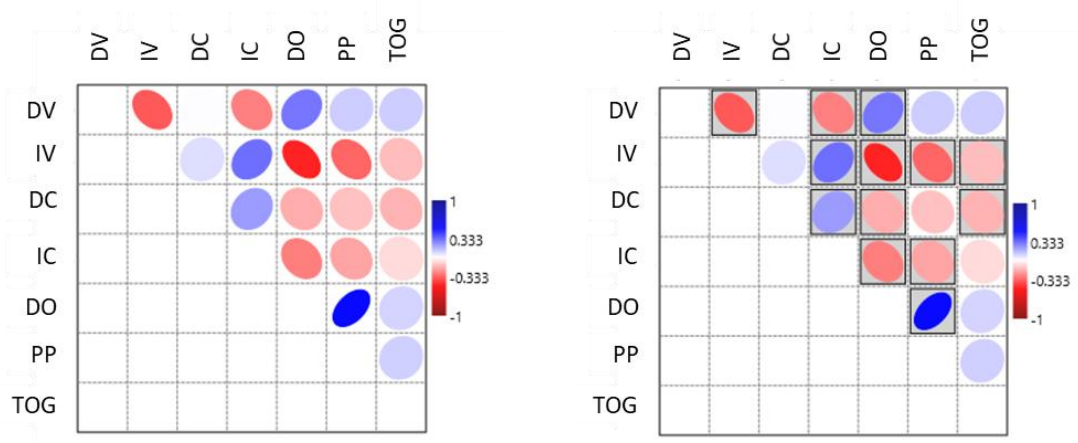
Figura 31. Curva ROC para o modelo RF

Em seguida, seguiu-se o fluxo de etapas apresentado na seção anterior. Considerando o conjunto de treinamento, é possível observar que as variáveis preditoras apresentam, ainda que moderadas, correlações significativas entre si, conforme pode ser visto na Tabela 9. Os valores superiores em representam os valores da correlação entre as variáveis e os valores inferiores são referentes ao *p-value* associado, sendo que valores de *p-value* inferiores a 0,05 indicam correlações significativas.

Tabela 9. Matriz de correlação

	IV	**DC**	**IC**	**DO**	**PP**	**TOG**
DV	-0,325 0,000	0,003 0,957	-0,247 0,000	0,267 0,000	0,098 0,109	0,097 0,110
IV		0,063 0,302	0,281 0,000	-0,430 0,000	-0,298 0,000	-0,128 0,035
DC			0,193 0,001	-0,158 0,009	-0,119 0,050	-0,146 0,016
IC				-0,251 0,000	-0,176 0,004	0,070 0,252
DO					0,482 0,000	0,083 0,176
PP						0,092 0,132

A Figura 32 ilustra os resultados numéricos apresentados na Tabela 9. As elipses com cores mais fortes indicam maiores valores de correlação. Caso a cor da elipse seja azul, a correlação é positiva e caso seja vermelha a correlação é negativa. Na Figura 32 ainda é possível distinguir entre as correlações significativas e não significativas, uma vez que aquelas que possuem *p-value* inferior a 0,05 foram inseridas em uma caixa cinza.



(a) (b)

Figura 32. (a) Gráfico de correlação; (b) Gráfico de correlação com destaque para as correlações significativas (*p-value* < 0,05)

A partir disso, conclui-se que existe a possibilidade de aplicação da análise fatorial. Entretanto, a fim de verificar se os dados são adequados para a realização desse procedimento, realizou-se o teste KMO disponível em linguagem R para esses dados, obtendo um valor de 0.70 para o overall, e individualmente todos os valores foram superiores a 0,6. A partir disso, conclui-se que os dados são adequados para a utilização análise fatorial.

Embora 5 fatores sejam suficientes para explicar mais que 84.0% da variância contida nos dados, optou-se por considerar 6 fatores, capazes de explicar 93.3% variabilidade presente nos dados, a fim de que não existam *loadings* com sinais opostos para as variáveis melhor explicadas por determinado fator *i*, uma vez que isso poderia prejudicar a interpretação da influência das variáveis na probabilidade de ocorrência de feição que será apresentada a seguir. A Tabela 10 mostra os valores de *loadings* e comunalidades resultantes da análise fatorial. Dessa forma a análise fatorial neste trabalho foi utilizada com o intuito de gerar novas variáveis, os *scores* dos fatores rotacionados, sem correlação, a fim de que métodos clássicos, como análise de um arranjo de superfície de resposta por

meio de mínimos quadrados ponderados pudesse ser utilizada, sem se preocupar com a matriz de variância e covariância entre as variáveis.

Tabela 10. *Loadings* dos fatores e comunalidades obtidas após análise fatorial

Variável	F1	F2	F3	F4	F5	F6	Comunalidade
PP	**0,924**	-0,036	0,04	0,000	-0,073	0,09	0,87
DO	**0,719**	-0,359	-0,228	0,147	0,026	0,061	0,724
IV	-0,204	**0,941**	0,14	-0,004	0,068	-0,133	0,969
DV	0,079	-0,14	**-0,971**	-0,02	-0,049	0,116	0,985
DC	-0,078	0,013	-0,018	**-0,985**	0,075	-0,093	0,992
TOG	0,044	-0,057	-0,046	0,073	**-0,991**	0,021	0,996
IC	-0,113	0,127	0,117	-0,098	0,022	**-0,972**	0,998
Var.	1,4393	1,0542	1,0322	1,0081	1,0019	0,9973	6,5329
% Var.	0,206	0,151	0,147	0,144	0,143	0,142	0,933

Em seguida, o banco de dados de treinamento foi dividido em duas partes de acordo com a ocorrência ou não de feição, ou seja, um conjunto de dados com todos os casos em que foram detectadas feições oleosas e outro em que para todos os casos não houve detecção de feições oleosas. Considerando o primeiro conjunto de dados de treinamento, constituído apenas de casos em que foi constatada a presença de feição oleosa, os *scores* dos fatores rotacionados foram convertidos em um arranjo de superfície de resposta. As probabilidades de ocorrência de feição obtidas anteriormente ao executar o método RF foram armazenadas como respostas desse arranjo. A partir disso, o modelo foi analisado por meio do algoritmo de mínimos quadrados ponderados (*Weighted Least Squares* – WLS), conforme utilizado em (LUZ *et al.*, 2021). O modelo encontrado para a probabilidade de ocorrência de feição pode ser vislumbrado pela Eq. (17) com valores de R^2 e $R^2_{ajustado}$ iguais a 98,60% e 98,35%, respectivamente.

$$\begin{aligned}\widehat{\boldsymbol{P}}_{\boldsymbol{RFC}} = {} & \mathbf{0,91835} + \mathbf{0,05445}\boldsymbol{F_1} - \mathbf{0,10487}\boldsymbol{F_2} - \mathbf{0,05100}\boldsymbol{F_3} - \mathbf{0,04612}\boldsymbol{F_4} \\ & - \mathbf{0,07502}\boldsymbol{F_5} + \mathbf{0,01667}\boldsymbol{F_6} - \mathbf{0,1126}\boldsymbol{F_2^2} - \mathbf{0,10042}\boldsymbol{F_3^2} \\ & - \mathbf{0,10987}\boldsymbol{F_6^2} + \mathbf{0,0296}\boldsymbol{F_1}\,\boldsymbol{F_2} - \mathbf{0,13547}\boldsymbol{F_2}\,\boldsymbol{F_3} + \mathbf{0,19621}\boldsymbol{F_3}\,\boldsymbol{F_6}\end{aligned} \quad (17)$$

Assim, a fim de obter uma maior compreensão dos efeitos associados a cada uma das variáveis utilizou-se o algoritmo *Desirability* disponível no software Minitab. Adotou-se um Target igual a 0,4 para a probabilidade de ocorrência de feição, e os fatores foram fixados de acordo com uma determinada observação com ocorrência de feição para o conjunto de treinamento, sendo que apenas um fator variou a cada rodada. O objetivo, nesse caso, foi compreender qual o impacto das variáveis sobre a probabilidade de

ocorrência de feições oleosas, conforme pode ser observado na Figura 33, e não de conseguir de fato reduzir a probabilidade para o Target informado.

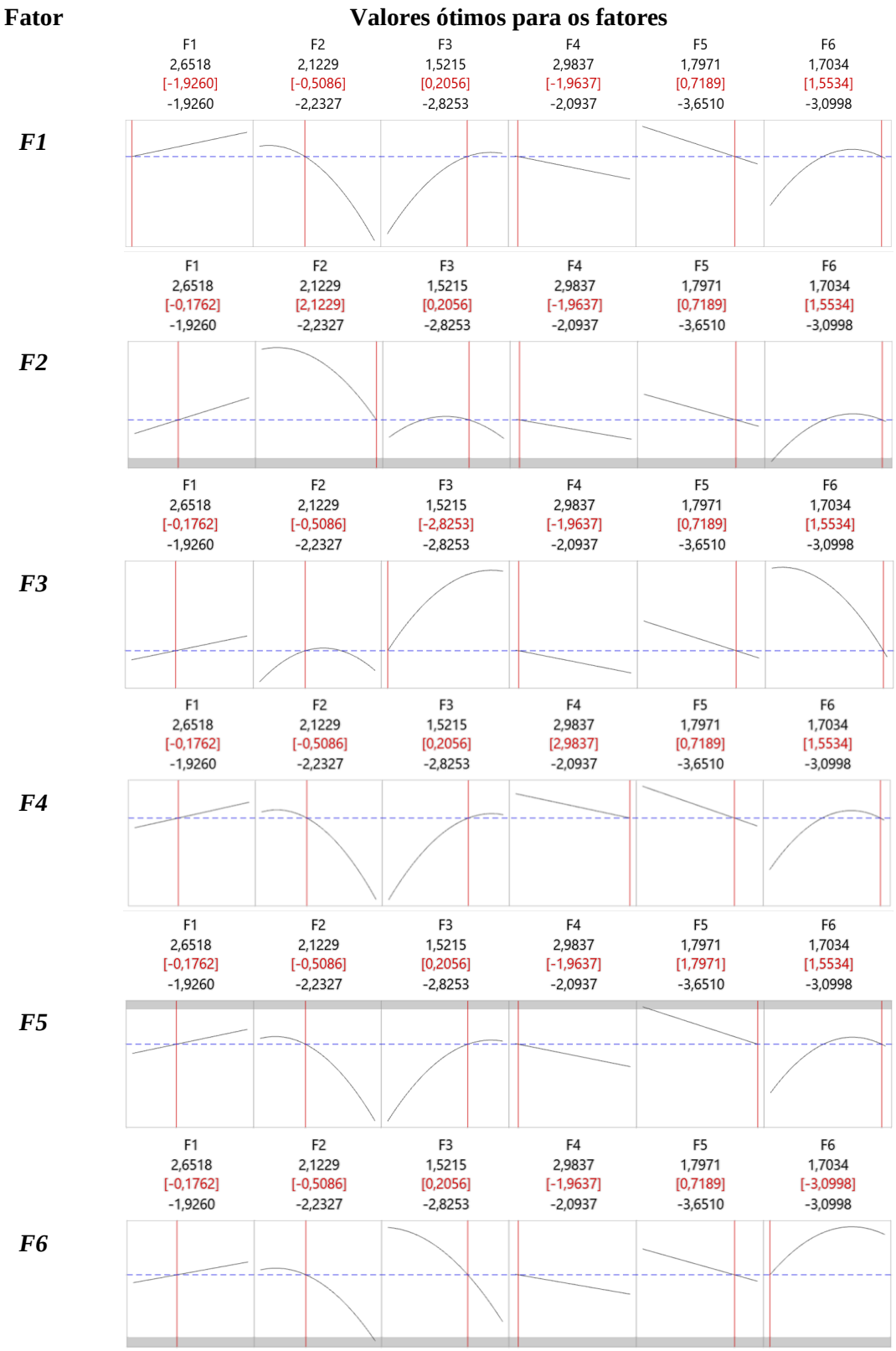


Figura 33. Resultados obtidos com o algoritmo *desirability*

Os gráficos de contorno e de superfície mostrados na Figura 34 mostram o comportamento da probabilidade considerando os fatores que contemplam a direção do vento e a intensidade da corrente enquanto as demais variáveis são mantidas constantes. É importante ressaltar que foram considerados em 3 cenários distintos: (a) nível de TOG igual ao valor de Q1 = 9,33333; (b) nível de TOG igual ao valor de Q2 = 11,8045 (mediana); (c) e nível de TOG igual ao valor de Q3 = 15,45835.

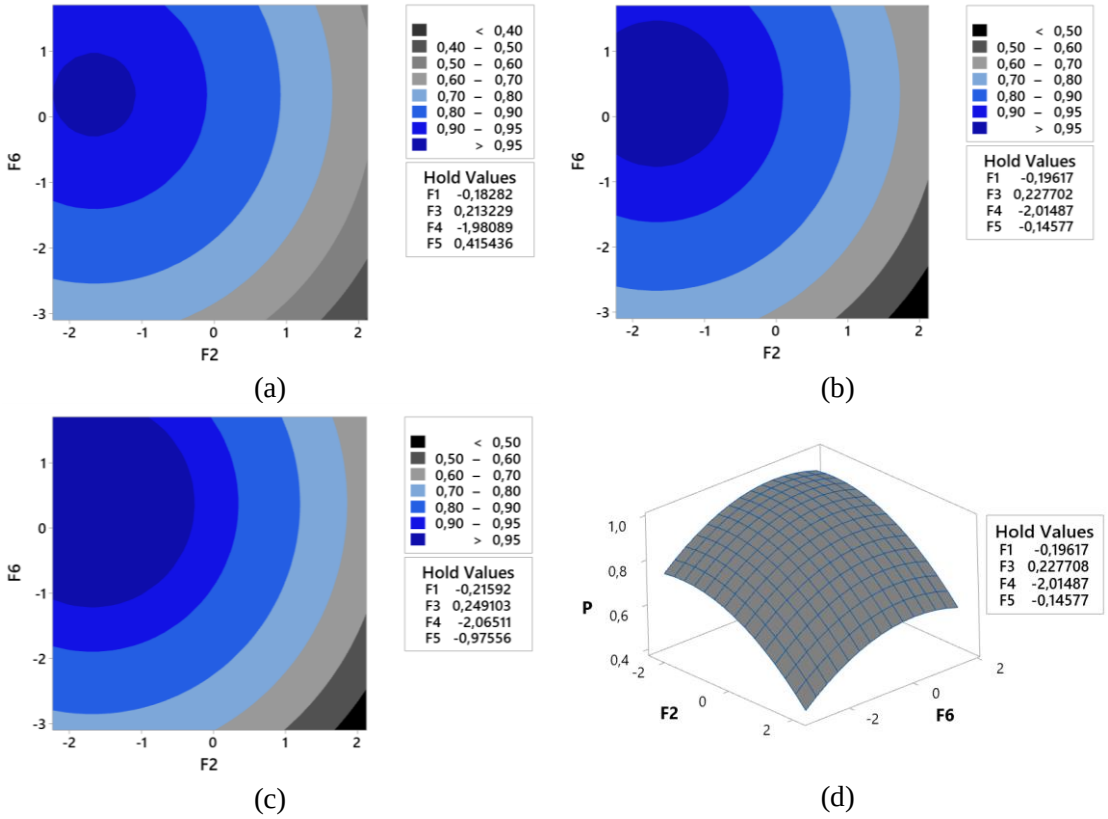


Figura 34. Gráficos de contorno e superfície para a o modelo da probabilidade de ocorrência e detecção de feições oleosas

É possível observar que quanto maior o valor do TOG Espectrofotométrico maior a área em azul escuro no centro dos gráficos de contorno, indicando a influência dessa variável na ocorrência e detecção de feições oleosas. Ao mesmo tempo pode-se observar que para maiores valores de F2, menor a probabilidade de ocorrência e detecção de feição, sendo que F2 está positivamente relacionado com a variável IV.

Em relação ao fator F6, quanto maior o seu valor, maior a probabilidade de ocorrência e detecção de feição. Convém destacar que F6 está negativamente correlacionado com a variável IC, dessa forma, menores valores de IC estão associados a maiores probabilidades, enquanto que maiores valores de IC estão relacionados maiores probabilidades de ocorrência e detecção de feições oleosas.

Os modelos das variáveis originais em função dos fatores podem ser observados nas equações de Eq. (18) até Eq. (24).

$$DV = 114{,}49 + 7{,}02\,F_1 - 9{,}86\,F_2 - 91{,}35\,F_3 - 1{,}656\,F_4 - 4{,}589\,F_5 + 10{,}83\,F_6 \quad (18)$$

$$IV = 5{,}3021 - 0{,}5024\,F_1 + 2{,}2457\,F_2 + 0{,}4079\,F_3 - 0{,}0182\,F_4 + 0{,}1734\,F_5 - 0{,}3361\,F_6 \quad (19)$$

$$DC = 186{,}878 - 4{,}447\,F_1 - 0{,}679\,F_2 - 0{,}386\,F_3 - 58{,}341\,F_4 + 4{,}453\,F_5 - 5{,}550\,F_6 \quad (20)$$

$$IC = 0{,}49233 - 0{,}03385\,F_1 + 0{,}04149\,F_2 + 0{,}03319\,F_3 - 0{,}02878\,F_4 + 0{,}00650\,F_5 - 0{,}28887\,F_6 \quad (21)$$

$$DO = 118{,}75 + 45{,}53\,F_1 - 31{,}25\,F_2 - 10{,}17\,F_3 + 8{,}53\,F_4 + 1{,}77\,F_5 + 3{,}62\,F_6 \quad (22)$$

$$PP = 10{,}576 + 2{,}2512\,F_1 + 0{,}155\,F_2 - 0{,}0102\,F_3 + 0{,}0162\,F_4 - 0{,}1826\,F_5 + 0{,}2263\,F_6 \quad (23)$$

$$TOG = 11{,}4212 + 0{,}2052\,F_1 - 0{,}3319\,F_2 - 0{,}1676\,F_3 + 0{,}3209\,F_4 - 4{,}4034\,F_5 + 0{,}0913\,F_6 \quad (24)$$

4.5. Modelagem da extensão da feição oleosa

Para avaliação do tamanho da feição oleosa, incialmente realizou-se o teste KMO para avaliar se os dados são propensos para a realização da técnica de análise fatorial. Nesse caso os dados ficaram no limite e muitas variáveis não apresentaram valores superior a 0,5. Dessa forma, optou-se por não se realizar a análise fatorial. É importante ressaltar que o banco de dados utilizado aqui é significativamente diferente daquele utilizado na seção 4.2, uma vez que os dados sem ocorrência de feição foram desconsiderados.

O modelo linear apresentou bons resultados e em virtude disso converteram-se os dados em um arranjo fatorial customizado. A Tabela 11 representa o modelo codificado obtido nessa etapa por meio do algoritmo de WLS, atingindo valores de R^2 e $R^2_{ajustado}$ iguais a 86,78% e 86,70%, respectivamente.

Conforme pode ser observado na Tabela 11, não há interações nem efeitos quadráticos considerados no modelo. Dessa forma, foi possível avaliar o gráfico de efeitos principais

demonstrado na Figura 35. É possível observar que quanto maiores os valores de variáveis como IC e TOG maior a extensão da feição e quanto maiores os valores de IV, PP e DV, menor a extensão da feição oleosa.

Tabela 11. Coeficientes para o modelo de extensão da feição

Termo	Efeito	*Coef*	*SE Coef*	*T-Value*	*P-Value*
Constante		3,818	0,103	37,24	0,000
DV	-1,655	-0,827	0,107	-7,72	0,000
IV	-5,234	-2,617	0,186	-14,04	0,000
IC	4,066	2,033	0,156	13,06	0,000
PP	-2,522	-1,261	0,139	-9,05	0,000
TOG	1,823	0,912	0,13	7,03	0,000

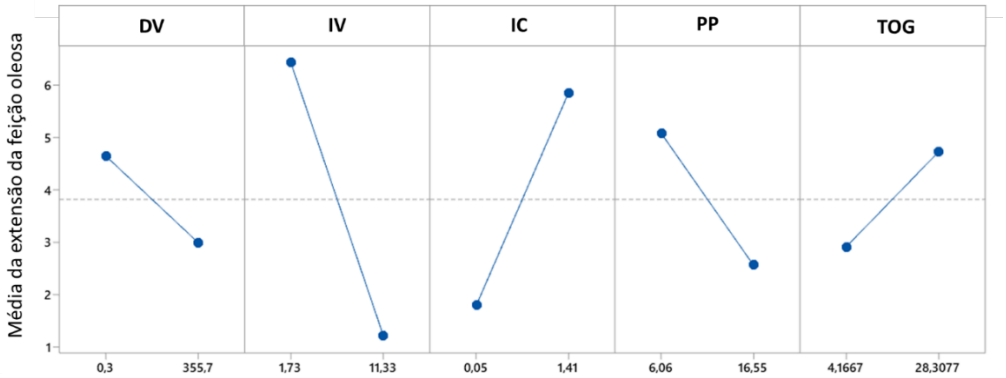


Figura 35. Gráfico de efeitos principais para os fatores do modelo de extensão da feição oleosa

O modelo de regressão para a extensão da feição oleosa (E_{fo}) em unidades decodificadas pode ser visualizado na Eq. (25).

$$E_{fo} = 7{,}516 - 0{,}004655\,WD - 0{,}5452\,WS + 2{,}990\,CS - 0{,}2404\,PP + 0{,}0755\,TOG \quad (25)$$

A fim de facilitar o entendimento dos efeitos das variáveis sobre a resposta considerada, ou seja, a extensão da feição oleosa, são apresentados na Figura 36 os gráficos de contorno considerando as variáveis intensidade do vento e intensidade da corrente para três diferentes níveis de TOG e a superfície de resposta associada.

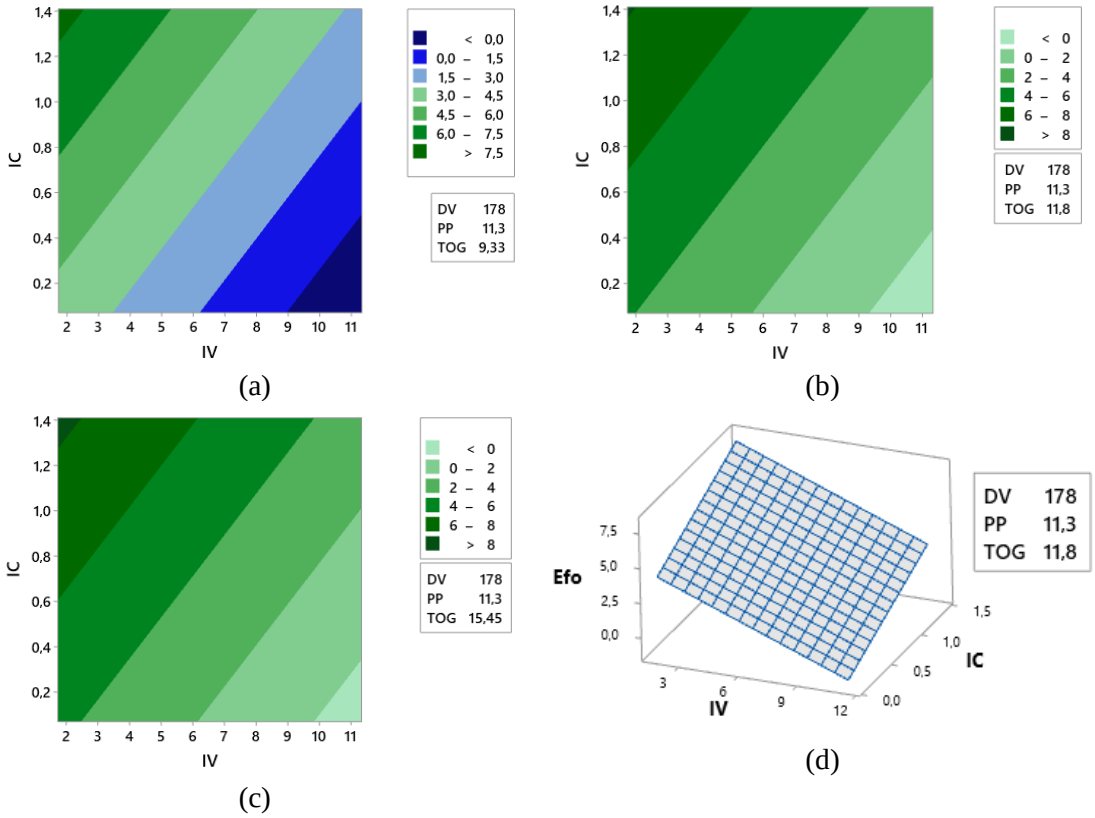


Figura 36. Gráficos de contorno e superfície para a o modelo da extensão de feições oleosas

Os níveis de TOG utilizados na Figura 36, foram os mesmos daqueles utilizados na Figura 34. Assim, é possível perceber as modificações nos gráficos de contorno à medida que o TOG Espectrofotométrico vai aumentando, ressaltando a importância dessa variável na extensão da feição oleosa, mesmo que não seja a medida oficialmente utilizada para fins fiscais. Além disso, é possível observar que a combinação de maiores valores da intensidade da corrente, juntamente com menores valores de intensidade do vento, estão associados a maiores valores de extensão da feição oleosa, conforme já observado na literatura.

5. CONCLUSÃO

O presente trabalho teve como objetivo principal estudar o impacto de variáveis meteo-oceanográficas e do total de óleos e graxas (TOG), obtido via espectrofotometria e absorção molecular, na probabilidade de formação e detecção de feições oleosas, bem como na extensão da feição. Os dados foram obtidos de uma plataforma *offshore* que realiza processamento primário de petróleo. Consideraram-se valores de direção do vento (DV), intensidade do vento (IV), direção da corrente (DC), intensidade da corrente (IC), direção da onda (DO), período de pico (PP) e TOG Espectrofotométrico.

Primeiramente, a partir do conjunto de dados disponibilizado, elaborou-se um arranjo fatorial completo com dois níveis para 6 variáveis meteo-oceanográficas e para o TOG. Primeiramente os valores do banco de dados foram codificados em -1 e $+1$, caso os valores das variáveis em cada uma das observações fossem menores ou maiores que os valores da mediana de cada uma delas, respectivamente. Foram, então, contabilizadas a quantidade de ocorrência e não ocorrência de feições oleosas para cada cenário, e consequentemente a probabilidade de ocorrência de feição para cada um dos 128 cenários. Além disso, foi obtida a probabilidade de ocorrência de cada um destes cenários.

O risco associado aos cenários foi obtido por meio da multiplicação da probabilidade de ocorrência de cada cenário pela probabilidade de ocorrência de feição oleosa. Os valores de riscos foram padronizados entre 0 e 1 a fim de facilitar a observação de cenários mais ou menos propensos a ocorrência de feições oleosas. Uma vez que foi possível observar que em determinados cenários, uma maior quantidade de casos, e consequentemente um maior risco associado, passaram-se às análises mais profundas para os dados considerados.

Realizou-se a comparação do desempenho de 5 métodos de classificação em relação ao problema binário de ocorrência de feição. Dentre os algoritmos considerados (*random forest* (RF), *k-nearest neighbors* (KNN), *multi-layer perceptron* (MLP), regressão logística binária (RLB) e *support vector machine* (SVM)) e também os *ensembles* dos 3 melhores classificadores (RF, RLB e SVM), o método RF apresentou melhores resultados em termos de acurácia. É importante destacar que o conjunto de dados foi divido em treinamento e teste de 50 formas distintas e para cada uma delas armazenaram-se as métricas de acurácia, especificidade e sensibilidade, de modo que a média de cada uma delas pudesse ser obtida e utilizada para a comparação.

Em seguida, o modelo construído utilizando o RF apresentou uma matriz de confusão satisfatória, bem como um elevado valor de área sob a chamada curva ROC (0,93). Dentre as variáveis que mais contribuíram para a separação entre as classes, a IV apresentou maior importância. Para identificar o efeito das variáveis de entrada, converteram-se os valores das variáveis preditoras em *scores* de fatores rotacionados que foram convertidos em um arranjo de superfície de resposta do tipo arranjo composto central (CCD). Os valores de probabilidade de ocorrência e detecção de feição foram armazenados e utilizados como resposta do arranjo. A técnica de *Desirability* foi utilizada demonstrando que quanto maiores os valores de IV, DV e IC menor será a probabilidade de ocorrência e detecção de feições oleosas, e quanto maiores os valores de TOG, PP, DO e DC maior essa probabilidade.

Valores de ventos mais intensos, conforme já observado na literatura, são menos propícios ao aparecimento de feições oleosas, da mesma forma que um valor de corrente mais intensa poderá aumentar a extensão da feição oleosa tornando-a mais fina e, portanto, de mais difícil detecção. Em relação ao TOG, espera-se que um maior valor de associado acarrete maior probabilidade de ocorrência e detecção de feição. Convém destacar que o TOG Espectrofotométrico, considerado nesse estudo, foi adequado para a previsão da feição oleosa, uma vez que se mostrou significativo durante a elaboração do modelo de probabilidade de detecção e ocorrência de feição oleosa.

Em relação ao período de pico, pode-se observar que um estado de mar com picos menores tem menor probabilidade de aparecimento de feições oleosas. Ao passo que valores maiores de período de pico aumentam a probabilidade do aparecimento de feições. Isso se explica pelo fato de um mar mais agitado ser menos propenso a formação de feições oleosas. A direção do vento exerceu influência, ainda que menor, no mesmo sentido da intensidade do vento em relação à probabilidade de ocorrência e detecção de feição, ao passo que a direção da onda e corrente exerceram influência no mesmo sentido do TOG.

Outra importante conclusão é que mesmo em casos que o TOG sendo baixo a depender das variáveis meteo-oceanográficas, ainda pode haver grande probabilidade de ocorrência de feição, conforme pode ser observado nos *boxplots* apresentados para as variáveis. Dessa forma, em conjunto com as informações obtidas pelo modelo, é possível concluir que as variáveis meteo-oceanográficas possuem influencia muito alta na

formação das feições oleosas e que nesses casos o valor do TOG deve ser o mínimo possível para reduzir ao máximo a probabilidade de ocorrência e detecção de feições.

Foi possível obter o modelo de extensão das feições oleosas a partir da conversão dos dados de ocorrência de feição em um arranjo do tipo fatorial. O gráfico de efeitos principais indicou que quanto maiores os valores de variáveis como IC e TOG maior a extensão da feição e quanto maiores os valores de IV, PP e DV, menor a extensão da feição oleosa. Dessa forma, constata-se que mares com correntes mais fortes, com períodos de picos menores, promovendo maior mistura podem contribuir para o aumento da extensão da feição oleosa. Dessa forma, a espessura associada é menor o que provavelmente poderá dificultar a detecção da feição, conforme resultado obtido no estágio anterior.

Além do exposto, convém ainda destacar que a metodologia proposta para a modelagem da probabilidade é adequada para utilização em conjuntos de dados quando se considera um cenário no qual existe pouco ou nenhum estudo prévio a respeito das influências das variáveis preditoras na resposta investigada, não se restringindo a apenas dados relacionados ao tema do presente estudo. Por fim, convém destacar que a clusterização dos dados com presença de feição para a criação do modelo foi determinante para encontrar um bom ajuste, uma vez que ao se considerar todos casos, com e sem feição, conjuntamente não foi possível obter um modelo que se ajustasse aos dados.

5.1. Contribuições do trabalho

Na literatura, são poucos os estudos que realizam uma análise de variáveis meteo-oceanográficas e de qualquer metodologia de medição do TOG no contexto de formação de feições oleosas. Poucas variáveis haviam sido consideradas em estudos já realizados, limitando-se na maioria das vezes a variáveis como intensidade do vento e intensidade da corrente.

Dessa forma, o presente estudo contribui principalmente no sentido de fornecer maior compreensão a respeito de quais variáveis são mais significativas para a probabilidade de ocorrência e detecção de feições oleosas, bem como na sua extensão. Além disso, a metodologia utilizada para compreensão desses efeitos é uma outra contribuição desse trabalho. Foram utilizadas técnicas multivariadas, planejamento de experimentos e *Desirability* de forma combinada, o que foi de extrema importância para a solução do problema.

Em termos mais práticos, o presente trabalho possui como implicação imediata a utilização do modelo de previsão da ocorrência e detecção de feições por plataformas *offshore* de processamento primário de petróleo. A partir da estimativa de um alto risco para detecção e ocorrência de feições oleosas, é possível operar de maneira a reduzir o nível do TOG a fim de que essa probabilidade também seja reduzida. Convém destacar que o nível do TOG adequado irá depender de cada cenário, desde que não ultrapasse os limites prá-estabelecidos, já que existe forte influência das variáveis meteo-oceanográficas.

5.2. Sugestões para trabalhos futuros

Como trabalhos futuros, sugere-se a aplicação de regressão logística para a classificação das observações relativas às feições oleosas, desta ou de outras plataformas *offshore* de processamento primário de petróleo. Sugerem-se, no mínimo, 30 conjuntos de treinamentos e teste distintos armazenado, então, o coeficiente associado às variáveis do modelo. Dessa forma, trabalha-se com coeficientes estocásticos que possuem média e variância associadas.

Uma outra maneira de melhorar a capacidade preditiva do classificador é realizar a sua combinação, entretanto com valores de pesos otimizados. É possível dividir o conjunto de dados em treinamento, validação e teste e a partir dos valores das métricas calculadas considerando o conjunto de validação é possível modelar tais métricas em função dos pesos e então otimizá-los a fim de fornecerem valores ótimos de acurácia, especificidade e sensibilidade. Aplicação de técnicas de planejamento de experimentos, como arranjos de misturas, podem ser muito úteis nesses casos.

Além disso, sugere-se considerar outras variáveis meteo-oceanográficas, como a altura significativa. Esta variável, pode exercer certa influência sobre a ocorrência de feições oleosas. No presente estudo, ela acabou sendo desconsiderada por conta da quantidade de dados faltantes. Também seria interessante englobar variáveis relacionadas ao processamento primário de petróleo, como por exemplo, pressão no flotador, vazão de saída, temperatura da água descartada, utilização de produtos químicos, entre outras.

6. REFERÊNCIAS

ABU-AISHEH, Zeina; RAVEAUX, Romain; RAMEL, Jean Yves. Efficient k-nearest neighbors search in graph space. *Pattern Recognition Letters*, v. 134, p. 77–86, 2020. Disponível em: <https://doi.org/10.1016/j.patrec.2018.05.001>.

AIZENBERG, Igor *et al.* Multilayer Neural Network with Multi-Valued Neurons in time series forecasting of oil production. *Neurocomputing*, v. 175, p. 980–989, 2016.

AL ADASANI, Ahmad; BAI, Baojun. Analysis of EOR projects and updated screening criteria. *Journal of Petroleum Science and Engineering*, v. 79, n. 1–2, p. 10–24, 2011.

BALESTRASSI, P. P. *et al.* Design of experiments on neural network's training for nonlinear time series forecasting. *Neurocomputing*, v. 72, n. 4–6, p. 1160–1178, 2009.

BALTHIS, William L. *et al.* Sediment quality benchmarks for assessing oil-related impacts to the deep-sea benthos. *Integrated Environmental Assessment and Management*, v. 13, n. 5, p. 840–851, 2017.

BASSIOUNI, Mahmoud M. *et al.* Intelligent hybrid approaches for human ECG signals identification. *Signal, Image and Video Processing*, v. 12, n. 5, p. 941–949, 2018. Disponível em: <https://doi.org/10.1007/s11760-018-1237-5>.

BELINATO, Gabriela *et al.* A multivariate normal boundary intersection PCA-based approach to reduce dimensionality in optimization problems for LBM process. *Engineering with Computers*, v. 35, n. 4, p. 1533–1544, 2019. Disponível em: <http://dx.doi.org/10.1007/s00366-018-0678-3>.

BERTRAND, J. Will M.; FRANSOO, Jan C. Operations management research methodologies using quantitative modeling. *International Journal of Operations and Production Management*, v. 22, n. 2, p. 241–264, 2002.

BISSACOT, A. C.G. *et al.* Comparison of neural networks and logistic regression in assessing the occurrence of failures in steel structures of transmission lines. *Open Electrical and Electronic Engineering Journal*, v. 10, p. 11–26, 2016.

BRANDÃO, Luiz Filipe Paiva; BRAGA, Jez Willian Batista; SUAREZ, Paulo Anselmo Ziani. Determination of vegetable oils and fats adulterants in diesel oil by high performance liquid chromatography and multivariate methods. *Journal of Chromatography A*, v. 1225, p. 150–157, 2012. Disponível em: <http://dx.doi.org/10.1016/j.chroma.2011.12.076>.

BREIMAN, L. Random Forests. *Machine Learning*, p. 5–32, 2001.

CARRANZA SÁNCHEZ, Yamid Alberto; DE OLIVEIRA, Silvio. Exergy analysis of offshore primary petroleum processing plant with CO2 capture. *Energy*, v. 88, p. 46–56, 2015.

CHAU, Ngoc Le; TRAN, Ngoc Thoai; DAO, Thanh-Phong. A hybrid approach of density-based topology, multilayer perceptron, and water cycle-moth flame algorithm for multi-stage optimal design of a flexure mechanism. *Engineering with Computers*, 2021. Disponível em: <https://doi.org/10.1007/s00366-021-01417-4>.

CHEN, Huazhou *et al.* A combination strategy of random forest and back propagation network for variable selection in spectral calibration. *Chemometrics and Intelligent*

Laboratory Systems, v. 182, n. February, p. 101–108, 2018. Disponível em: <https://doi.org/10.1016/j.chemolab.2018.09.002>.

CHEN, Yifu *et al.* A data-driven binary-classification framework for oil fingerprinting analysis. *Environmental Research*, v. 201, n. March, 2021.

CONAMA Resolution No. 393/2007. Disponível em: <http://www.braziliannr.com/brazilian-envi%0Aronmentallegislation/conama-resolution-39307/>.

CUI, Ying *et al.* Classification of estrogen receptor selective compounds with benzopyranskeleton using counterpropagation artificial neural networks optimised by genetic algorithms. *Chemometrics and Intelligent Laboratory Systems*, v. 146, p. 385–395, 2015. Disponível em: <http://dx.doi.org/10.1016/j.chemolab.2015.06.007>.

DAI, Bo *et al.* Statistical model optimized random forest regression model for concrete dam deformation monitoring. *Structural Control and Health Monitoring*, v. 25, n. 6, p. 1–15, 2018.

DANESHGAR ASL, Samira *et al.* Hindcast modeling of oil slick persistence from natural seeps. *Remote Sensing of Environment*, v. 189, p. 96–107, 2017. Disponível em: <http://dx.doi.org/10.1016/j.rse.2016.11.003>.

DE ALMEIDA, Fabricio Alves *et al.* A gage study through the weighting of latent variables under orthogonal rotation. *IEEE Access*, v. 8, p. 183557–183570, 2020.

DEKAMIN, Azam *et al.* FIUS: Fixed partitioning undersampling method. *Clinica Chimica Acta*, v. 522, n. April, p. 174–183, 2021.

DING, Jun; BAR-JOSEPH, Ziv. MethRaFo: MeDIP-seq methylation estimate using a Random Forest Regressor. *Bioinformatics (Oxford, England)*, v. 33, n. 21, p. 3477–3479, 2017.

EL-DAHSHAN, El Sayed A.; BASSIOUNI, Mahmoud M. Computational intelligence techniques for human brain MRI classification. *International Journal of Imaging Systems and Technology*, v. 28, n. 2, p. 132–148, 2018.

ERIŞTI, Hüseyin; UÇAR, Ayşegül; DEMIR, Yakup. Wavelet-based feature extraction and selection for classification of power system disturbances using support vector machines. *Electric Power Systems Research*, v. 80, n. 7, p. 743–752, 2010.

FÁVERO, L. P. *Análise de dados: Técnicas Multivariadas Exploratórias com SPSS e STATA*. [S.l.]: Elsevier, Grupo Gen, 2015.

FAWCETT, Tom. An introduction to ROC analysis. *Pattern Recognition Letters*, v. 27, n. 8, p. 861–874, 2006.

FU, Guang Hui *et al.* Stable biomarker screening and classification by subsampling-based sparse regularization coupled with support vector machines in metabolomics. *Chemometrics and Intelligent Laboratory Systems*, v. 160, n. November 2016, p. 22–31, 2017.

GANBOLD, G; CHASIA, & S. Comparison between Possibilistic c-Means (PCM) and Artificial Neural Network (ANN) Classification Algorithms in Land use/ Land cover Classification. *International Journal of Knowledge Content Development & Technology*, v. 7, n. 1, p. 57–78, 2017. Disponível em:

<http://dx.doi.org/10.5865/IJKCT.2017.7.1.057>.

GARCIA-PINEDA, Oscar *et al.* Classification of oil spill by thicknesses using multiple remote sensors. *Remote Sensing of Environment*, v. 236, n. August 2019, 2020.

GHADERYAN, Peyvand; ABBASI, Ataollah; SEDAAGHI, Mohammad Hossein. An efficient seizure prediction method using KNN-based undersampling and linear frequency measures. *Journal of Neuroscience Methods*, v. 232, p. 134–142, 2014. Disponível em: <http://dx.doi.org/10.1016/j.jneumeth.2014.05.019>.

GHORBANZAD'E, Mehdi; FATEMI, Mohammad Hossein. Classification of central nervous system agents by least squares support vector machine based on their structural descriptors: A comparative study. *Chemometrics and Intelligent Laboratory Systems*, v. 110, n. 1, p. 102–107, 2012. Disponível em: <http://dx.doi.org/10.1016/j.chemolab.2011.10.003>.

HAMEED, Mohammed Majeed *et al.* Prediction of high-strength concrete: high-order response surface methodology modeling approach. *Engineering with Computers*, 2021. Disponível em: <https://doi.org/10.1007/s00366-021-01284-z>.

HASSANAT, Ahmad Basheer *et al. Solving the Problem of the K Parameter in the KNN Classifier Using an Ensemble Learning Approach*. . [S.l: s.n.]. Disponível em: <http://arxiv.org/abs/1409.0919>. , 2014

HAYKIN, S. *Neural Networks and Learning Machines*. 3rd editio ed. [S.l.]: Prentice Hall, 2009.

HE, Ting *et al.* The detonation heat prediction of nitrogen-containing compounds based on quantitative structure-activity relationship (QSAR) combined with random forest (RF). *Chemometrics and Intelligent Laboratory Systems*, v. 213, n. September 2020, p. 104249, 2021. Disponível em: <https://doi.org/10.1016/j.chemolab.2021.104249>.

HEGDE, Chiranth; MILLWATER, Harry; GRAY, Ken. Classification of drilling stick slip severity using machine learning. *Journal of Petroleum Science and Engineering*, v. 179, n. January, p. 1023–1036, 2019.

HOSMER JR., DAVID W.; LEMESHOW, STANLEY; STURDIVANT, RODNEY X. *Applied Logistic Regression*. Third Edit ed. Hoboken, New Jersey: John Wiley & Sons, Inc., 2013.

HSU, Chang S.; ROBINSON, Paul R. *Practical Advances in Petroleum Processing Volume 1*. New York: Springer Science + Business Media Inc., 2006.

ISLAM, Md Shafiqul *et al.* Solid waste bin detection and classification using Dynamic Time Warping and MLP classifier. *Waste Management*, v. 34, n. 2, p. 281–290, 2014.

JAVADI, Arash *et al.* A combination of artificial neural network and genetic algorithm to optimize gas injection: A case study for EOR applications. *Journal of Molecular Liquids*, v. 339, p. 116654, 2021. Disponível em: <https://doi.org/10.1016/j.molliq.2021.116654>.

JIMÉNEZ-CORDERO, Asunción; MORALES, Juan Miguel; PINEDA, Salvador. A novel embedded min-max approach for feature selection in nonlinear Support Vector Machine classification. *European Journal of Operational Research*, v. 293, n. 1, p. 24–35, 2021.

JOHNSON, Richard A.; WICHERN, Dean W. *Applied multivariate statistical analysis*. Sixth ed. Upper Saddle River, New Jersey: Pearson Education, Inc., 2007.

JONES, David S. J. “STAN”; PUJADÓ, Peter R. *Handbook of Petroleum Processing*. Dordrecht: Springer, 2006.

KLEMZ, Ana Caroline *et al.* Treatment of real oilfield produced water by liquid-liquid extraction and efficient phase separation in a mixer-settler based on phase inversion. *Chemical Engineering Journal*, v. 417, n. November 2020, 2021.

KOZIARSKI, Michał. Radial-Based Undersampling for imbalanced data classification. *Pattern Recognition*, v. 102, 2020.

KUO, Hung Fei; FARICHA, Anifatul. Artificial Neural Network for Diffraction Based Overlay Measurement. *IEEE Access*, v. 4, n. 104, p. 7479–7486, 2016.

LAZRI, Mourad; AMEUR, Soltane. Combination of support vector machine, artificial neural network and random forest for improving the classification of convective and stratiform rain using spectral features of SEVIRI data. *Atmospheric Research*, v. 203, n. November 2017, p. 118–129, 2018. Disponível em: <https://doi.org/10.1016/j.atmosres.2017.12.006>.

LEWIS, Alun. Current Status of the Baoac (Bonn Agreement Oil Appearance Code). n. January, p. 19, 2007.

LI, Guofa *et al.* Tool wear state recognition based on gradient boosting decision tree and hybrid classification RBM. *International Journal of Advanced Manufacturing Technology*, v. 110, n. 1–2, p. 511–522, 2020.

LI, Hongdong; LIANG, Yizeng; XU, Qingsong. Support vector machines and its applications in chemistry. *Chemometrics and Intelligent Laboratory Systems*, v. 95, n. 2, p. 188–198, 2009. Disponível em: <http://dx.doi.org/10.1016/j.chemolab.2008.10.007>.

LIANG, Jing *et al.* Data fusion of laser induced breakdown spectroscopy (LIBS) and infrared spectroscopy (IR) coupled with random forest (RF) for the classification and discrimination of compound salvia miltiorrhiza. *Chemometrics and Intelligent Laboratory Systems*, v. 207, n. September, p. 104179, 2020. Disponível em: <https://doi.org/10.1016/j.chemolab.2020.104179>.

LIN, Shun Ku *et al.* Classification of patients with Alzheimer ’ s disease using the arterial pulse spectrum and a multilayer - perceptron analysis. p. 1–14, 2021.

LIN, Wei Chao *et al.* Clustering-based undersampling in class-imbalanced data. *Information Sciences*, v. 409–410, p. 17–26, 2017.

LIU, Fang; ZHOU, Zhiguang. A new data classification method based on chaotic particle swarm optimization and least square-support vector machine. *Chemometrics and Intelligent Laboratory Systems*, v. 147, p. 147–156, 2015. Disponível em: <http://dx.doi.org/10.1016/j.chemolab.2015.08.015>.

LIU, Xu Ying; WU, Jianxin; ZHOU, Zhi Hua. Exploratory undersampling for class-imbalance learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, v. 39, n. 2, p. 539–550, 2009.

LUO, Dan *et al.* Secondary oil recovery using graphene-based amphiphilic Janus nanosheet fluid at an ultralow concentration. *Industrial and Engineering Chemistry*

Research, v. 56, n. 39, p. 11125–11132, 2017.

LUZ, E.R. *et al.* A new multiobjective optimization with elliptical constraints approach for nonlinear models implemented in a stainless steel cladding process. *International Journal of Advanced Manufacturing Technology*, 2021.

MANGILI, Ivan *et al.* Modeling and optimization of ultrasonic devulcanization using the response surface methodology based on central composite face-centered design. *Chemometrics and Intelligent Laboratory Systems*, v. 144, p. 1–10, 2015. Disponível em: <http://dx.doi.org/10.1016/j.chemolab.2015.03.003>.

MANSOURI, Kamel *et al.* Quantitative structure-activity relationship models for ready biodegradability of chemicals. *Journal of Chemical Information and Modeling*, v. 53, n. 4, p. 867–878, 2013.

MARINS, Matheus A. *et al.* Fault detection and classification in oil wells and production/service lines using random forest. *Journal of Petroleum Science and Engineering*, v. 197, n. March 2020, p. 107879, 2021. Disponível em: <https://doi.org/10.1016/j.petrol.2020.107879>.

MEHDIZADEH, Saeid; FATHIAN, Farshad; ADAMOWSKI, Jan F. Hybrid artificial intelligence-time series models for monthly streamflow modeling. *Applied Soft Computing Journal*, v. 80, p. 873–887, 2019. Disponível em: <https://doi.org/10.1016/j.asoc.2019.03.046>.

MITICHE, Imene *et al.* Classification of EMI discharge sources using time–frequency features and multi-class support vector machine. *Electric Power Systems Research*, v. 163, n. February, p. 261–269, 2018. Disponível em: <https://doi.org/10.1016/j.epsr.2018.06.016>.

MOGENSEN, Kristian; MASALMEH, Shehadeh. A review of EOR techniques for carbonate reservoirs in challenging geological settings. *Journal of Petroleum Science and Engineering*, v. 195, n. May, p. 107889, 2020. Disponível em: <https://doi.org/10.1016/j.petrol.2020.107889>.

MONTGOMERY, Douglas C. *Design and Analysis of Experiments*. 9. ed. New York: John Wiley & Sons, 2017.

MORAIS, Camilo L.M.; LIMA, Kássio M.G. Comparing unfolded and two-dimensional discriminant analysis and support vector machines for classification of EEM data. *Chemometrics and Intelligent Laboratory Systems*, v. 170, n. June, p. 1–12, 2017.

MORANDIN, Lora A.; O'HARA, Patrick D. Offshore oil and gas, and operational sheen occurrence: Is there potential harm to marine birds? *Environmental Reviews*, v. 24, n. 3, p. 285–318, 2016.

MÜLLER, Andreas C.; GUIDO, Sarah. *Introduction to Machine Learning with Python a guide for datat scientists*. 1st. ed. Gravenstein Highway North, Sebastopol: O'Reilly Media, Inc., 2016.

MUSBAH, Hmeda; ALY, Hamed H.; LITTLE, Timothy A. Energy management of hybrid energy system sources based on machine learning classification algorithms. *Electric Power Systems Research*, v. 199, n. June, p. 107436, 2021. Disponível em: <https://doi.org/10.1016/j.epsr.2021.107436>.

MYERS, Raymond H.; MONTGOMERY, Douglas C.; ANDERSON-COOK, Christine M. *Response Surface Methodology Process and product optimization using designed experimetns*. Fourth ed. Honoken, New Jersey: John Wiley & Sons, Inc., 2016.

NATIONAL OCEANIC AND ATMOSPHERIC ADMINISTRATION - NOAA. *Open Water Oil Identification Job Aid (NOAA-CODE) for Aerial Observation. With Standardized Oil Slick Appearence and Structure Nomeclature and Codes Rep. .* Seattle, Washington: [s.n.]. , 2016

NEFF, Jerry; LEE, Kenneth; DEBLOIS, Elisabeth M. Produced Water: Overview of Composition, Fates, and Effects. In: LEE, KENNETH; NEFF, JERRY (Org.). . *Produced Water*. New York, NY: Springer New York, 2011. v. 1. p. 3–54. Disponível em: <http://link.springer.com/10.1007/978-1-4614-0046-2_1>.

O'HARA, Patrick D.; MORANDIN, Lora A. Effects of sheens associated with offshore oil and gas development on the feather microstructure of pelagic seabirds. *Marine Pollution Bulletin*, v. 60, n. 5, p. 672–678, 2010. Disponível em: <http://dx.doi.org/10.1016/j.marpolbul.2009.12.008>.

OLOSO, Munirudeen A. *et al.* Ensemble SVM for characterisation of crude oil viscosity. *Journal of Petroleum Exploration and Production Technology*, v. 8, n. 2, p. 531–546, 2018.

OLSON, Jacob; VALOVA, Iren; MICHEL, Howard. WSCISOM: wireless sensor data cluster identification through a hybrid SOM/MLP/RBF architecture. *Progress in Artificial Intelligence*, v. 5, n. 4, p. 233–250, 2016.

OTCHERE, Daniel Asante *et al.* Application of supervised machine learning paradigms in the prediction of petroleum reservoir properties: Comparative analysis of ANN and SVM models. *Journal of Petroleum Science and Engineering*, v. 200, n. August 2020, p. 108182, 2021. Disponível em: <https://doi.org/10.1016/j.petrol.2020.108182>.

PALANISAMY, K.; MURUGAPPAN, M.; YAACOB, S. Multiple Physiological Signal-Based Human Stress Identification Using Non-Linear Classifiers. *Electronics and Electrical Engineering*, v. 19, n. 7, 11 set. 2013. Disponível em: <http://www.eejournal.ktu.lt/index.php/elt/article/view/2232>.

PERES, António M. *et al.* Chemometric classification of several olive cultivars from Trás-os-Montes region (northeast of Portugal) using artificial neural networks. *Chemometrics and Intelligent Laboratory Systems*, v. 105, n. 1, p. 65–73, 2011. Disponível em: <http://dx.doi.org/10.1016/j.chemolab.2010.11.001>.

PISANO, A. *et al.* An oceanographic survey for oil spill monitoring and model forecasting validation using remote sensing and in situ data in the Mediterranean Sea. *Deep-Sea Research Part II: Topical Studies in Oceanography*, v. 133, p. 132–145, 2016.

PUGGINA BIANCHESI, N.M. *et al.* A design of experiments comparative study on clustering methods. *IEEE Access*, v. 7, 2019.

QUINTANILHA, Igor M. *et al.* A fault detector/classifier for closed-ring power generators using machine learning. *Reliability Engineering & System Safety*, v. 212, n. March, p. 107614, 2021. Disponível em: <https://doi.org/10.1016/j.ress.2021.107614>.

RAI, Praveen; LONDHE, Narendra D.; RAJ, Ritesh. Fault classification in power

system distribution network integrated with distributed generators using CNN. *Electric Power Systems Research*, v. 192, n. September 2020, p. 106914, 2021. Disponível em: <https://doi.org/10.1016/j.epsr.2020.106914>.

RAMAMURTHY, Arun *et al.* ML-Based Classification of Device Environment Using Wi-Fi and Cellular Signal Measurements. *IEEE Access*, v. 10, p. 29461–29472, 2022.

RIBEIRO, J. S. *et al.* Simultaneous optimization of the microextraction of coffee volatiles using response surface methodology and principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, v. 102, n. 1, p. 45–52, 2010.

ROCHA, Luiz Célio Souza *et al.* Toward a robust optimal point selection: a multiple-criteria decision-making process applied to multi-objective optimization using response surface methodology. *Engineering with Computers*, 2020. Disponível em: <https://doi.org/10.1007/s00366-020-00973-5>.

SACHANIYA, Bhumi K. *et al.* Bioengineering for multiple PAHs degradation for contaminated sediments: Response surface methodology (RSM) and artificial neural network (ANN). *Chemometrics and Intelligent Laboratory Systems*, v. 202, n. May, p. 104033, 2020. Disponível em: <https://doi.org/10.1016/j.chemolab.2020.104033>.

SALEM, Shiva; JAFARZADEH-GHOUSHCHI, Saeid. Estimation of optimal physico-chemical characteristics of nano-sized inorganic blue pigment by combined artificial neural network and response surface methodology. *Chemometrics and Intelligent Laboratory Systems*, v. 159, n. October, p. 80–88, 2016.

SANQUETTA, Carlos R. *et al.* Volume estimation of Cryptomeria japonica logs in southern Brazil using artificial intelligence models. *Southern Forests*, v. 80, n. 1, p. 29–36, 2018.

SARAVANA KUMAR, Ramachandran; MANIKANDAN, Parasuraman. Medical big data classification using a combination of random forest classifier and K-means clustering. *International Journal of Intelligent Systems and Applications*, v. 10, n. 11, p. 11–19, 2018.

SAVOLAINEN, M. A.; KAZMIERCZAK, R. F.; CAFFEY, R. H. Determining the effect of environmental accidents on responses to a Gulf of Mexico recreational for-hire fishing industry surveya. *Journal of Fish Biology*, v. 83, n. 4, p. 1035–1045, 2013.

SILVA, Ivan Nunes Da *et al. Artificial Neural Networks a practical course*. Switzerland: Springer International Publishing AG Switzerland, 2017.

SOLEYMANI, Fazlollah; MASNAVI, Houman; SHATEYI, Stanford. Classifying a lending portfolio of loans with dynamic updates via a machine learning technique. *Mathematics*, v. 9, n. 1, p. 1–15, 2021.

SPEISER, Jaime Lynn *et al.* BiMM forest: A random forest method for modeling clustered and longitudinal binary outcomes. *Chemometrics and Intelligent Laboratory Systems*, v. 185, n. January, p. 122–134, 2019. Disponível em: <https://doi.org/10.1016/j.chemolab.2019.01.002>.

THAI, Duc Kien *et al.* Gradient tree boosting machine learning on predicting the failure modes of the RC panels under impact loads. *Engineering with Computers*, v. 37, n. 1, p. 597–608, 2021. Disponível em: <https://doi.org/10.1007/s00366-019-00842-w>.

THURSTONE, L.L. *Multiple-factor analysis*. Chicago: University of Chicago Press.,

1947.

TRIGGIA, Attílio Alberto *et al. Fundamentos de Engenharia de Petróleo*. Rio de Janeiro: Editora Interciência Ltda., 2001.

VIACAVA, Gabriela Elena; ROURA, Sara Inés; AGÜERO, María Victoria. Optimization of critical parameters during antioxidants extraction from butterhead lettuce to simultaneously enhance polyphenols and antioxidant activity. *Chemometrics and Intelligent Laboratory Systems*, v. 146, p. 47–54, 2015. Disponível em: <http://dx.doi.org/10.1016/j.chemolab.2015.05.002>.

WAN, Shaohua *et al.* Deep Multi-Layer perceptron classifier for behavior analysis to estimate Parkinson's disease severity using smartphones. *IEEE Access*, v. 6, p. 36825–36833, 2018.

WANG, Hong; MOAYEDI, Hossein; KOK FOONG, Loke. Genetic algorithm hybridized with multilayer perceptron to have an economical slope stability design. *Engineering with Computers*, 2020. Disponível em: <https://doi.org/10.1007/s00366-020-00957-5>.

WMO. *Guide to Instruments and Methods of Observation Volume I - Measurement of Meteorological Variables*. . Geneva: World Meteorological Organization. , 2018

YANG, Ming. Measurement of Oil in Produced Water. *Produced Water*. New York, NY: Springer New York, 2011. p. 57–88. Disponível em: <http://link.springer.com/10.1007/978-1-4614-0046-2_2>.

YEAP, Danny *et al.* Peak detection and random forests classification software for gas chromatography/differential mobility spectrometry (GC/DMS) data. *Chemometrics and Intelligent Laboratory Systems*, v. 203, n. April, p. 104085, 2020. Disponível em: <https://doi.org/10.1016/j.chemolab.2020.104085>.

YELIPE, Usha Rani; PORIKA, Sammulal; GOLLA, Madhu. An efficient approach for imputation and classification of medical data values using class-based clustering of medical records. *Computers and Electrical Engineering*, v. 66, p. 487–504, 2018.

YILMAZ, Iik; KAYNAR, Oguz. Multiple regression, ANN (RBF, MLP) and ANFIS models for prediction of swell potential of clayey soils. *Expert Systems with Applications*, v. 38, n. 5, p. 5958–5966, 2011.

YONG, Hua Hie *et al.* Identifying smoker subgroups with high versus low smoking cessation attempt probability: A decision tree analysis approach. *Addictive Behaviors*, v. 103, n. November 2019, p. 106258, 2020. Disponível em: <https://doi.org/10.1016/j.addbeh.2019.106258>.

YU, Hualong; NI, Jun; ZHAO, Jing. ACOSampling: An ant colony optimization-based undersampling method for classifying imbalanced DNA microarray data. *Neurocomputing*, v. 101, p. 309–318, 2013. Disponível em: <http://dx.doi.org/10.1016/j.neucom.2012.08.018>.

YU, Keqiang *et al.* Response surface methodology for optimizing LIBS testing parameters: A case to conduct the elemental contents analysis in soil. *Chemometrics and Intelligent Laboratory Systems*, v. 195, n. October, p. 103891, 2019. Disponível em: <https://doi.org/10.1016/j.chemolab.2019.103891>.

YU, Xinliang; YU, Yixiong; ZENG, Qun. Support vector machine classification of

streptavidin-binding aptamers. *PLoS ONE*, v. 9, n. 6, p. 1–5, 2014.

ZADKARAMI, Morteza; SHAHBAZIAN, Mehdi; SALAHSHOOR, Karim. Pipeline leakage detection and isolation: An integrated approach of statistical and wavelet feature extraction with multi-layer perceptron neural network (MLPNN). *Journal of Loss Prevention in the Process Industries*, v. 43, p. 479–487, 2016. Disponível em: <http://dx.doi.org/10.1016/j.jlp.2016.06.018>.

ZATSEPA, S. N. *et al.* The Role of Wind Waves in Oil Spill Natural Dispersion in the Sea. *Oceanology*, v. 58, n. 4, p. 517–524, 2018.

ZHANG, Shichao; CHENG, Debo; *et al.* A novel kNN algorithm with data-driven k parameter computation. *Pattern Recognition Letters*, v. 109, p. 44–54, 2018. Disponível em: <https://doi.org/10.1016/j.patrec.2017.09.036>.

ZHANG, Shichao; LI, Xuelong; *et al.* Efficient kNN classification with different numbers of nearest neighbors. *IEEE Transactions on Neural Networks and Learning Systems*, v. 29, n. 5, p. 1774–1785, 2018.

ZHANG, Shichao. Nearest neighbor selection for iteratively kNN imputation. *Journal of Systems and Software*, v. 85, n. 11, p. 2541–2552, 2012. Disponível em: <http://dx.doi.org/10.1016/j.jss.2012.05.073>.

ZHANG, Tianlong *et al.* Classification of steel samples by laser-induced breakdown spectroscopy and random forest. *Chemometrics and Intelligent Laboratory Systems*, v. 157, p. 196–201, 2016. Disponível em: <http://dx.doi.org/10.1016/j.chemolab.2016.07.001>.

ZHU, Shan *et al.* Equivalent circuit model recognition of electrochemical impedance spectroscopy via machine learning. *Journal of Electroanalytical Chemistry*, v. 855, n. August, p. 113627, 2019. Disponível em: <https://doi.org/10.1016/j.jelechem.2019.113627>.

ZUO, Wangmeng; ZHANG, David; WANG, Kuanquan. On kernel difference-weighted k-nearest neighbor classification. *Pattern Analysis and Applications*, v. 11, n. 3–4, p. 247–257, 2008.

APÊNDICE

A presente seção tem como objetivo apresentar os resultados para os testes t pareado desenvolvidos a fim de comparar o desempenho dos classificadores individuais com o *ensemble* de classificadores (RF, RLB e SVM). As Figuras Ap 1, Ap 2 e Ap 3 representam os testes relacionados a comparação do desempenho do *ensemble* 1 e as Figuras Ap 4, AP 5 e Ap 6 estão relacionadas ao *ensemble* 2.

Estatísticas Descritivas

Amostra	N	Média	DesvPad	EP Média
Ac_ENS	50	0,7773	0,0812	0,0115
Ac_RF	50	0,7840	0,0760	0,0107

Estimativa da diferença pareada

Média	DesvPad	EP Média	IC de 95% da diferença_μ
-0,00667	0,04949	0,00700	(-0,02073; 0,00740)

diferença_μ: média de (Ac_ENS - Ac_RF)

Teste

Hipótese nula	H_0: diferença_μ = 0
Hipótese alternativa	H_1: diferença_μ ≠ 0

Valor-T	Valor-p
-0,95	0,345

Estatísticas Descritivas

Amostra	N	Média	DesvPad	EP Média
Ac_ENS	50	0,7773	0,0812	0,0115
Ac_RLB	50	0,7473	0,0886	0,0125

Estimativa da diferença pareada

Média	DesvPad	EP Média	IC de 95% da diferença_μ
0,03000	0,06108	0,00864	(0,01264; 0,04736)

diferença_μ: média de (Ac_ENS - Ac_RLB)

Teste

Hipótese nula	H_0: diferença_μ = 0
Hipótese alternativa	H_1: diferença_μ ≠ 0

Valor-T	Valor-p
3,47	0,001

Estatísticas Descritivas

Amostra	N	Média	DesvPad	EP Média
Ac_ENS	50	0,7773	0,0812	0,0115
Ac_SVM	50	0,7680	0,0747	0,0106

Estimativa da diferença pareada

Média	DesvPad	EP Média	IC de 95% da diferença_μ
0,00933	0,05130	0,00726	(-0,00525; 0,02391)

diferença_μ: média de (Ac_ENS - Ac_SVM)

Teste

Hipótese nula	H_0: diferença_μ = 0
Hipótese alternativa	H_1: diferença_μ ≠ 0

Valor-T	Valor-p
1,29	0,204

Figura Ap 1. Teste t pareado para a Acurácia em relação ao *ensemble* 1

Estatísticas Descritivas

Amostra	N	Média	DesvPad	EP Média
Sp_ENS	50	0,7091	0,1179	0,0167
Sp_RF	50	0,7473	0,1105	0,0156

Estimativa da diferença pareada

Média	DesvPad	EP Média	IC de 95% da diferença_μ
-0,0382	0,0776	0,0110	(-0,0602; -0,0161)

diferença_μ: média de (Sp_ENS - Sp_RF)

Teste

Hipótese nula H_0: diferença_μ = 0
Hipótese alternativa H_1: diferença_μ ≠ 0

Valor-T	Valor-p
-3,48	0,001

Estatísticas Descritivas

Amostra	N	Média	DesvPad	EP Média
Sp_ENS	50	0,7091	0,1179	0,0167
Sp_RLB	50	0,7192	0,1066	0,0151

Estimativa da diferença pareada

Média	DesvPad	EP Média	IC de 95% da diferença_μ
-0,0101	0,0851	0,0120	(-0,0343; 0,0141)

diferença_μ: média de (Sp_ENS - Sp_RLB)

Teste

Hipótese nula H_0: diferença_μ = 0
Hipótese alternativa H_1: diferença_μ ≠ 0

Valor-T	Valor-p
-0,84	0,404

Estatísticas Descritivas

Amostra	N	Média	DesvPad	EP Média
Sp_ENS	50	0,7091	0,1179	0,0167
Sp_SVM	50	0,7018	0,1127	0,0159

Estimativa da diferença pareada

Média	DesvPad	EP Média	IC de 95% da diferença_μ
0,00726	0,06047	0,00855	(-0,00993; 0,02444)

diferença_μ: média de (Sp_ENS - Sp_SVM)

Teste

Hipótese nula H_0: diferença_μ = 0
Hipótese alternativa H_1: diferença_μ ≠ 0

Valor-T	Valor-p
0,85	0,400

Figura Ap 2. Teste t pareado para a Especificidade em relação ao *ensemble* 1

Estatísticas Descritivas

Amostra	N	Média	DesvPad	EP Média
Sn_ENS	50	0,8316	0,0975	0,0138
Sn_RF	50	0,8122	0,1030	0,0146

Estimativa da diferença pareada

Média	DesvPad	EP Média	IC de 95% da diferença_μ
0,01938	0,06912	0,00978	(-0,00026; 0,03903)

diferença_μ: média de (Sn_ENS - Sn_RF)

Teste

Hipótese nula	H_0: diferença_μ = 0
Hipótese alternativa	H_1: diferença_μ ≠ 0

Valor-T	Valor-p
1,98	0,053

Estatísticas Descritivas

Amostra	N	Média	DesvPad	EP Média
Sn_ENS	50	0,8316	0,0975	0,0138
Sn_RLB	50	0,7676	0,1263	0,0179

Estimativa da diferença pareada

Média	DesvPad	EP Média	IC de 95% da diferença_μ
0,0640	0,0821	0,0116	(0,0407; 0,0873)

diferença_μ: média de (Sn_ENS - Sn_RLB)

Teste

Hipótese nula	H_0: diferença_μ = 0
Hipótese alternativa	H_1: diferença_μ ≠ 0

Valor-T	Valor-p
5,51	0,000

Estatísticas Descritivas

Amostra	N	Média	DesvPad	EP Média
Sn_ENS	50	0,8316	0,0975	0,0138
Sn_SVM	50	0,8237	0,0959	0,0136

Estimativa da diferença pareada

Média	DesvPad	EP Média	IC de 95% da diferença_μ
0,0079	0,0774	0,0109	(-0,0141; 0,0299)

diferença_μ: média de (Sn_ENS - Sn_SVM)

Teste

Hipótese nula	H_0: diferença_μ = 0
Hipótese alternativa	H_1: diferença_μ ≠ 0

Valor-T	Valor-p
0,72	0,474

Figura Ap 3. Teste t pareado para a Sensibilidade em relação ao *ensemble* 1

Estatísticas Descritivas

Amostra	N	Média	DesvPad	EP Média
Ac_ENS	50	0,7713	0,0825	0,0117
Ac_RF	50	0,7840	0,0760	0,0107

Estimativa da diferença pareada

Média	DesvPad	EP Média	IC de 95% da diferença_μ
-0,01267	0,04410	0,00624	(-0,02520; -0,00013)

diferença_μ: média de (Ac_ENS - Ac_RF)

Teste

Hipótese nula	H_0: diferença_μ = 0
Hipótese alternativa	H_1: diferença_μ ≠ 0

Valor-T	Valor-p
-2,03	0,048

Estatísticas Descritivas

Amostra	N	Média	DesvPad	EP Média
Ac_ENS	50	0,7713	0,0825	0,0117
Ac_RL	50	0,7473	0,0886	0,0125

Estimativa da diferença pareada

Média	DesvPad	EP Média	IC de 95% da diferença_μ
0,02400	0,05304	0,00750	(0,00893; 0,03907)

diferença_μ: média de (Ac_ENS - Ac_RL)

Teste

Hipótese nula	H_0: diferença_μ = 0
Hipótese alternativa	H_1: diferença_μ ≠ 0

Valor-T	Valor-p
3,20	0,002

Estatísticas Descritivas

Amostra	N	Média	DesvPad	EP Média
Ac_ENS	50	0,7713	0,0825	0,0117
Ac_SVM	50	0,7680	0,0747	0,0106

Estimativa da diferença pareada

Média	DesvPad	EP Média	IC de 95% da diferença_μ
0,00333	0,05184	0,00733	(-0,01140; 0,01807)

diferença_μ: média de (Ac_ENS - Ac_SVM)

Teste

Hipótese nula	H_0: diferença_μ = 0
Hipótese alternativa	H_1: diferença_μ ≠ 0

Valor-T	Valor-p
0,45	0,651

Figura Ap 4. Teste t pareado para a Sensibilidade em relação ao *ensemble* 2

Estatísticas Descritivas

Amostra	N	Média	DesvPad	EP Média
Sp_ENS	50	0,7002	0,1165	0,0165
Sp_RF	50	0,7473	0,1105	0,0156

Estimativa da diferença pareada

Média	DesvPad	EP Média	IC de 95% da diferença_μ
-0,0470	0,0726	0,0103	(-0,0677; -0,0264)

diferença_μ: média de (Sp_ENS - Sp_RF)

Teste

Hipótese nula	H_0: diferença_μ = 0
Hipótese alternativa	H_1: diferença_μ ≠ 0

Valor-T	Valor-p
-4,58	0,000

Estatísticas Descritivas

Amostra	N	Média	DesvPad	EP Média
Sp_ENS	50	0,7002	0,1165	0,0165
Sp_RL	50	0,7192	0,1066	0,0151

Estimativa da diferença pareada

Média	DesvPad	EP Média	IC de 95% da diferença_μ
-0,0190	0,0829	0,0117	(-0,0426; 0,0046)

diferença_μ: média de (Sp_ENS - Sp_RL)

Teste

Hipótese nula	H_0: diferença_μ = 0
Hipótese alternativa	H_1: diferença_μ ≠ 0

Valor-T	Valor-p
-1,62	0,112

Estatísticas Descritivas

Amostra	N	Média	DesvPad	EP Média
Sp_ENS	50	0,7002	0,1165	0,0165
Sp_SVM	50	0,7018	0,1127	0,0159

Estimativa da diferença pareada

Média	DesvPad	EP Média	IC de 95% da diferença_μ
-0,00161	0,05360	0,00758	(-0,01684; 0,01363)

diferença_μ: média de (Sp_ENS - Sp_SVM)

Teste

Hipótese nula	H_0: diferença_μ = 0
Hipótese alternativa	H_1: diferença_μ ≠ 0

Valor-T	Valor-p
-0,21	0,833

Figura Ap 5. Teste t pareado para a Sensibilidade em relação ao *ensemble* 2

Estatísticas Descritivas

Amostra	N	Média	DesvPad	EP Média
Sn_ENS	50	0,8274	0,0974	0,0138
Sn_RF	50	0,8122	0,1030	0,0146

Estimativa da diferença pareada

Média	DesvPad	EP Média	IC de 95% da diferença_μ
0,01514	0,06036	0,00854	(-0,00201; 0,03230)

diferença_μ: média de (Sn_ENS - Sn_RF)

Teste

Hipótese nula	H_0: diferença_μ = 0
Hipótese alternativa	H_1: diferença_μ ≠ 0

Valor-T	Valor-p
1,77	0,082

Estatísticas Descritivas

Amostra	N	Média	DesvPad	EP Média
Sn_ENS	50	0,8274	0,0974	0,0138
Sn_RL	50	0,7676	0,1263	0,0179

Estimativa da diferença pareada

Média	DesvPad	EP Média	IC de 95% da diferença_μ
0,0598	0,0780	0,0110	(0,0376; 0,0819)

diferença_μ: média de (Sn_ENS - Sn_RL)

Teste

Hipótese nula	H_0: diferença_μ = 0
Hipótese alternativa	H_1: diferença_μ ≠ 0

Valor-T	Valor-p
5,42	0,000

Estatísticas Descritivas

Amostra	N	Média	DesvPad	EP Média
Sn_ENS	50	0,8274	0,0974	0,0138
Sn_SVM	50	0,8237	0,0959	0,0136

Estimativa da diferença pareada

Média	DesvPad	EP Média	IC de 95% da diferença_μ
0,0037	0,0761	0,0108	(-0,0180; 0,0253)

diferença_μ: média de (Sn_ENS - Sn_SVM)

Teste

Hipótese nula	H_0: diferença_μ = 0
Hipótese alternativa	H_1: diferença_μ ≠ 0

Valor-T	Valor-p
0,34	0,736

Figura Ap 6. Teste t pareado para a Sensibilidade em relação ao ensemble 2

ANEXO A

Os dados utilizados no presente trabalho são apresentados na Tabela A 1 do presente anexo. É importante destacar que as células preenchidas com '*nan*' representam valores faltantes na base dados e, como mencionado anteriormente, foram substituídos pela média de suas respectivas colunas. Já a Tabela A 2 indica os dados utilizados para a modelagem da extensão da feição oleosa.

Tabela A 1. Conjunto de dados usados no presente estudo

Data e hora	***DV* (°)**	***IV* *(m/s)***	***DC* (°)**	***IC* *(m/s)***	***DO* (°)**	***PP* *(s)***	***TOG* *(mg/L)***	***Feição***
18/04/2018 05:10	113,20	6,55	196,24	0,71	158,19	7,96	14,17	não
02/05/2018 05:02	235,10	5,22	230,73	0,40	173,79	11,63	8,50	não
04/05/2018 18:06	63,20	5,57	258,26	0,31	158,27	12,76	12,08	não
09/05/2018 05:19	127,20	9,36	252,14	0,40	153,23	7,56	16,00	não
11/05/2018 05:07	72,80	3,38	216,01	0,49	153,85	10,76	7,75	não
12/05/2018 05:05	351,00	8,87	213,77	0,55	63,40	nan	8,08	não
14/05/2018 18:15	114,60	4,94	231,16	0,46	210,09	15,39	7,83	não
17/05/2018 05:19	212,30	3,97	89,29	0,11	90,67	12,49	13,69	não
21/05/2018 18:10	201,70	3,40	357,89	0,26	215,28	12,70	20,86	não
22/05/2018 17:34	17,40	3,23	8,15	0,09	189,05	11,88	19,21	não
28/05/2018 18:06	111,20	7,37	257,92	0,18	162,46	12,81	16,83	não
05/06/2018 05:05	152,70	4,49	199,89	0,04	209,67	10,76	13,58	não
05/06/2018 05:10	152,70	4,49	199,89	0,04	209,67	10,76	13,58	não
14/06/2018 18:10	156,20	10,98	259,94	0,38	188,91	14,47	11,17	não
01/07/2018 18:15	8,80	3,02	257,59	0,22	77,79	9,02	17,58	sim
15/07/2018 16:58	10,90	4,98	203,91	0,12	0,00	7,57	10,75	sim
18/07/2018 05:55	41,40	5,47	199,18	0,85	162,19	11,56	13,75	não

Continuação da Tabela A 1

Data e hora	***DV* (°)**	***IV* (m/s)**	***DC* (°)**	***IC* (m/s)**	***DO* (°)**	***PP* (s)**	***TOG* (mg/L)**	***Feição***
18/07/2018 17:04	33,30	4,19	187,68	1,00	170,46	11,53	13,75	sim
23/07/2018 05:10	40,40	3,14	202,30	1,03	208,80	10,45	16,85	não
25/07/2018 18:15	nan	nan	nan	nan	153,25	18,04	10,08	sim
28/07/2018 16:52	36,50	5,41	200,34	0,61	128,61	7,89	5,42	não
01/08/2018 18:10	137,80	5,23	173,25	0,27	185,25	11,33	19,31	não
02/08/2018 16:46	118,80	8,50	166,24	0,30	186,53	11,61	15,00	não
02/08/2018 17:41	103,40	8,32	160,70	0,27	178,02	12,64	15,00	não
03/08/2018 05:55	92,70	7,85	160,06	0,36	155,08	12,13	15,33	não
05/08/2018 05:19	241,60	2,31	199,17	0,19	125,26	8,12	17,77	sim
06/08/2018 16:52	248,20	3,12	56,98	0,01	149,49	9,03	10,50	não
07/08/2018 05:07	180,90	4,63	180,58	0,16	172,83	8,83	11,00	não
08/08/2018 05:49	156,80	3,66	185,05	0,33	183,29	9,17	11,00	não
08/08/2018 16:58	89,30	4,04	158,46	0,51	153,97	9,69	11,00	não
08/08/2018 18:06	74,40	3,68	178,39	0,61	155,70	8,66	11,00	não
10/08/2018 05:13	3,80	13,80	192,58	1,04	19,85	7,44	5,58	não
12/08/2018 05:49	239,10	10,03	187,74	0,30	237,96	12,80	5,67	não
13/08/2018 17:40	214,20	7,89	58,09	0,12	229,54	8,35	3,75	não
15/08/2018 05:55	55,20	1,73	165,70	0,26	nan	nan	4,58	sim
15/08/2018 17:04	113,00	1,53	180,34	0,17	170,35	9,16	4,58	não
16/08/2018 05:05	35,50	1,07	187,47	0,17	148,78	8,70	5,58	não
17/08/2018 16:58	132,60	3,43	191,17	0,49	206,97	12,26	6,50	sim
18/08/2018 05:13	174,80	6,19	203,92	0,93	204,84	11,46	7,75	não
19/08/2018 05:55	96,30	2,95	203,29	0,71	179,33	8,07	6,83	não

Continuação da Tabela A 1

Data e hora	***DV* (°)**	***IV* (m/s)**	***DC* (°)**	***IC* (m/s)**	***DO* (°)**	***PP* (s)**	***TOG* (mg/L)**	***Feição***
24/08/2018 16:58	36,10	10,66	187,26	0,60	79,02	4,95	13,00	não
26/08/2018 05:13	20,50	8,51	218,78	0,82	62,87	7,83	10,25	não
26/08/2018 16:46	256,60	4,50	nan	nan	102,85	13,33	10,25	sim
29/08/2018 05:19	108,60	6,08	206,86	0,61	158,04	11,45	9,67	não
30/08/2018 16:46	45,10	4,74	208,74	0,98	113,01	9,16	11,08	não
02/09/2018 16:58	27,90	11,25	195,72	1,45	40,32	6,38	3,50	não
03/09/2018 05:13	18,20	12,93	188,16	1,29	47,11	6,65	4,08	não
04/09/2018 18:19	216,30	8,64	187,62	0,12	128,83	8,01	7,92	não
06/09/2018 05:49	201,20	8,00	200,88	0,66	229,86	9,43	4,08	não
09/09/2018 05:49	14,20	1,93	210,94	0,95	159,25	14,14	4,83	não
09/09/2018 16:57	97,50	2,54	214,20	0,72	146,80	14,10	4,83	sim
11/09/2018 17:34	150,00	4,00	211,07	0,53	172,27	10,86	5,45	sim
11/09/2018 18:14	150,00	4,00	211,07	0,53	172,27	10,86	5,45	sim
12/09/2018 06:01	148,00	3,82	212,05	0,56	182,30	9,10	3,86	não
14/09/2018 16:52	39,90	9,92	196,38	0,59	49,79	5,36	7,17	não
15/09/2018 16:46	358,20	8,12	201,80	0,48	37,87	7,09	13,17	não
17/09/2018 05:55	262,80	1,81	195,71	0,54	182,19	7,46	7,67	sim
19/09/2018 16:46	12,70	6,73	205,19	0,88	100,11	12,28	8,67	não
21/09/2018 05:05	39,90	11,18	208,13	0,98	72,35	10,71	8,33	não
21/09/2018 05:49	31,80	12,49	208,07	1,00	66,12	10,55	8,33	não
24/09/2018 17:28	45,80	6,31	196,73	1,10	85,19	10,61	8,50	não
25/09/2018 05:25	31,10	7,61	201,73	0,72	51,01	10,61	8,11	não
25/09/2018 18:06	71,20	8,35	204,64	0,73	59,92	11,28	8,11	não

Continuação da Tabela A 1

Data e hora	***DV* (°)**	***IV* (m/s)**	***DC* (°)**	***IC* (m/s)**	***DO* (°)**	***PP* (s)**	***TOG* (mg/L)**	***Feição***
30/09/2018 16:52	60,20	4,06	208,83	0,49	155,10	10,07	5,50	sim
01/10/2018 16:46	41,40	9,13	193,28	0,74	110,25	10,49	4,83	não
02/10/2018 05:55	16,40	10,70	190,81	0,65	45,42	6,12	13,33	não
02/10/2018 17:04	32,60	11,91	204,22	0,80	37,61	7,13	13,33	não
03/10/2018 05:10	25,50	11,38	211,75	0,92	29,53	7,56	13,58	não
03/10/2018 05:25	25,50	11,38	211,75	0,92	29,53	7,56	13,58	não
04/10/2018 16:58	242,90	2,39	203,01	0,92	113,18	8,70	11,92	sim
05/10/2018 05:13	192,60	5,68	213,24	1,09	nan	10,65	8,00	não
05/10/2018 06:01	185,90	7,34	207,86	1,08	179,66	9,87	8,00	não
06/10/2018 05:55	163,60	3,87	202,92	0,71	190,77	9,20	4,17	sim
08/10/2018 05:19	52,10	9,22	239,53	0,67	107,80	14,18	8,50	não
11/10/2018 05:49	6,50	5,61	206,02	1,12	71,98	7,12	10,69	não
11/10/2018 16:58	9,90	4,16	211,21	0,99	68,36	7,69	10,69	não
11/10/2018 17:46	354,70	2,65	211,51	1,09	75,86	8,43	10,69	sim
13/10/2018 16:46	21,50	11,29	187,81	1,02	45,72	13,90	10,92	não
15/10/2018 05:49	179,90	2,72	171,18	0,62	207,57	12,13	11,08	sim
16/10/2018 16:52	102,90	8,26	187,71	0,31	167,17	7,91	11,25	não
19/10/2018 05:25	16,70	5,41	197,43	1,00	100,00	14,46	14,75	não
21/10/2018 17:34	160,80	4,49	211,17	0,63	190,03	9,03	21,23	não
22/10/2018 17:04	78,90	5,40	206,37	0,91	184,05	11,67	17,00	não
24/10/2018 17:40	109,80	3,95	188,11	0,74	145,21	10,46	15,33	não
01/11/2018 05:19	39,10	8,87	198,62	0,83	140,43	12,79	11,33	não
01/11/2018 16:52	10,10	7,31	203,16	0,49	113,55	11,05	11,33	não

Continuação da Tabela A 1

Data e hora	***DV* (°)**	***IV* (m/s)**	***DC* (°)**	***IC* (m/s)**	***DO* (°)**	***PP* (s)**	***TOG* (mg/L)**	***Feição***
02/11/2018 16:46	172,90	5,75	246,59	0,56	158,96	10,87	12,00	não
03/11/2018 05:55	75,50	3,36	209,24	0,60	150,69	11,47	10,67	sim
08/11/2018 05:05	126,40	7,66	244,51	0,45	163,77	9,67	7,00	não
12/11/2018 17:57	93,40	6,56	223,11	0,29	209,95	8,14	14,31	sim
14/11/2018 05:13	42,60	8,33	206,78	0,42	81,19	6,07	13,17	não
14/11/2018 16:46	49,40	9,98	203,36	0,39	62,87	6,92	13,17	não
16/11/2018 18:04	214,40	3,22	222,49	0,29	75,71	6,28	17,50	sim
18/11/2018 16:46	31,70	10,78	203,77	0,53	48,51	6,12	16,17	não
20/11/2018 18:33	163,70	3,85	228,72	0,39	210,51	9,00	10,67	sim
20/11/2018 18:46	163,70	3,85	228,72	0,39	210,51	9,00	10,67	sim
22/11/2018 18:14	59,30	8,79	239,03	0,36	134,41	13,93	10,33	não
24/11/2018 05:49	37,00	9,68	222,19	0,46	54,54	6,17	11,33	não
25/11/2018 05:19	34,80	10,03	220,25	0,36	55,09	6,96	14,33	não
25/11/2018 05:49	32,50	9,67	229,50	0,40	49,07	7,87	14,33	não
26/11/2018 19:52	72,30	5,14	240,37	0,41	75,11	7,59	12,75	sim
27/11/2018 06:55	255,20	2,40	237,19	0,23	185,17	8,01	11,92	sim
27/11/2018 17:29	163,90	4,64	294,28	0,14	180,35	7,96	11,92	não
28/11/2018 06:49	106,60	1,93	239,19	0,28	166,01	13,41	8,67	sim
29/11/2018 06:07	43,40	4,46	251,97	0,20	123,07	8,06	7,67	não
02/12/2018 05:49	272,60	3,94	232,92	0,32	39,31	8,05	11,17	não
02/12/2018 06:04	272,60	3,94	232,92	0,32	39,31	8,05	11,17	sim
02/12/2018 06:49	269,10	3,62	235,84	0,30	49,52	7,81	11,17	sim
04/12/2018 06:01	156,90	3,71	205,44	0,33	208,68	9,07	9,75	não

Continuação da Tabela A 1

Data e hora	***DV* (°)**	***IV* (m/s)**	***DC* (°)**	***IC* (m/s)**	***DO* (°)**	***PP* (s)**	***TOG* (mg/L)**	***Feição***
04/12/2018 07:01	145,30	4,32	205,73	0,27	205,65	9,35	9,75	sim
06/12/2018 18:06	175,60	6,36	266,89	0,25	219,57	11,68	7,42	não
09/12/2018 18:19	110,50	5,66	303,81	0,12	179,30	13,20	8,92	não
10/12/2018 05:49	104,20	6,54	300,67	0,12	185,35	10,81	8,42	não
11/12/2018 06:19	34,30	5,00	227,02	0,23	123,07	10,68	9,92	sim
13/12/2018 17:28	26,80	8,04	205,29	0,50	43,47	7,04	10,42	não
14/12/2018 11:43	13,10	9,39	226,97	0,66	38,16	6,06	8,67	sim
14/12/2018 16:58	31,60	7,46	203,21	0,76	40,51	8,05	8,67	não
15/12/2018 06:07	34,90	6,95	197,80	0,94	39,60	6,74	8,83	não
16/12/2018 05:13	48,10	8,45	203,59	0,89	63,98	5,82	9,00	não
19/12/2018 05:19	27,20	7,61	224,21	0,51	33,65	6,29	8,17	não
21/12/2018 17:28	39,30	9,64	217,11	0,80	10,34	14,11	6,50	não
22/12/2018 05:25	32,00	8,69	218,67	0,60	21,00	6,74	4,42	não
22/12/2018 16:58	34,50	8,38	216,09	0,68	15,05	5,83	4,42	não
25/12/2018 05:55	239,10	4,78	240,65	0,43	54,50	7,07	5,00	não
25/12/2018 06:55	216,80	4,90	239,97	0,47	81,11	7,11	5,00	sim
26/12/2018 10:50	183,50	3,60	208,57	0,46	nan	nan	5,50	sim
27/12/2018 06:19	183,20	2,06	nan	nan	nan	nan	7,25	sim
29/12/2018 06:55	332,60	6,23	236,07	0,39	34,17	6,83	10,00	sim
30/12/2018 16:58	45,30	4,28	219,00	0,44	51,83	7,95	13,25	não
30/12/2018 18:46	55,40	5,20	216,33	0,49	47,30	7,20	13,25	sim
30/12/2018 19:07	55,40	5,20	216,33	0,49	47,30	7,20	13,25	sim
31/12/2018 07:07	2,10	11,33	221,47	0,45	20,67	7,08	12,25	sim

Continuação da Tabela A 1

Data e hora	***DV* (°)**	***IV* (m/s)**	***DC* (°)**	***IC* (m/s)**	***DO* (°)**	***PP* (s)**	***TOG* (mg/L)**	***Feição***
01/01/2019 16:46	39,40	9,28	217,65	0,85	27,29	6,12	13,08	não
04/01/2019 05:19	2,00	10,04	206,20	0,78	17,01	7,16	10,25	não
05/01/2019 06:01	4,30	9,33	225,08	0,52	26,24	7,12	11,42	não
06/01/2019 05:55	13,80	8,44	227,15	0,54	22,45	6,51	16,33	não
07/01/2019 05:05	24,20	8,67	221,71	0,48	23,81	6,06	19,67	não
09/01/2019 16:46	56,10	5,30	190,26	1,19	57,26	6,95	13,75	não
19/01/2019 05:05	29,30	9,46	222,34	0,66	10,85	6,18	9,67	não
20/01/2019 16:52	37,80	9,25	212,59	0,97	35,80	6,74	20,33	não
21/01/2019 05:01	22,70	9,59	227,89	0,57	23,78	11,90	8,25	não
23/01/2019 06:25	41,70	5,82	234,78	0,65	nan	12,72	6,75	sim
24/01/2019 17:37	50,10	8,45	200,60	0,99	71,74	7,17	17,17	não
27/01/2019 05:49	55,50	5,83	205,14	1,08	122,07	11,10	7,75	não
28/01/2019 05:19	54,40	6,91	199,82	0,85	133,82	7,67	8,50	não
29/01/2019 06:07	44,50	4,26	210,46	0,87	95,96	7,03	10,33	não
29/01/2019 16:52	60,70	5,19	210,12	0,87	101,99	7,25	10,33	não
30/01/2019 05:55	32,60	6,85	229,04	0,63	75,72	8,50	8,92	não
30/01/2019 17:28	39,00	6,15	194,67	0,83	86,71	7,30	8,92	não
31/01/2019 05:05	23,60	8,11	219,78	0,66	33,61	4,94	6,08	não
04/02/2019 17:04	183,30	7,37	215,72	0,99	nan	11,82	13,67	não
07/02/2019 06:06	290,20	4,16	339,55	0,15	nan	nan	14,00	sim
07/02/2019 06:55	303,70	3,49	299,53	0,06	203,31	12,18	14,00	sim
10/02/2019 05:13	26,80	3,98	308,39	0,08	81,36	12,31	5,00	não
10/02/2019 06:01	14,10	3,21	84,36	0,08	103,80	13,57	5,00	não

Continuação da Tabela A 1

Data e hora	***DV* (°)**	***IV* (m/s)**	***DC* (°)**	***IC* (m/s)**	***DO* (°)**	***PP* (s)**	***TOG* (mg/L)**	***Feição***
11/02/2019 05:55	3,20	10,86	168,96	1,00	49,86	6,31	8,00	não
12/02/2019 05:05	14,40	8,86	190,93	1,17	35,25	13,09	11,33	não
15/02/2019 17:28	133,20	5,50	121,19	0,31	159,60	7,55	5,83	não
18/02/2019 17:34	17,70	4,24	205,98	1,28	123,57	13,61	7,42	não
19/02/2019 17:04	4,10	5,06	26,33	0,05	179,45	12,88	4,75	não
19/02/2019 18:18	8,80	5,45	204,67	0,14	198,19	12,81	4,75	sim
20/02/2019 05:49	354,50	6,58	112,51	0,17	217,04	12,00	6,25	não
20/02/2019 17:21	16,80	3,50	123,52	0,16	169,70	12,81	6,25	sim
21/02/2019 16:52	241,30	3,89	148,68	0,28	208,00	12,69	8,67	não
26/02/2019 05:13	11,40	4,30	181,18	0,49	342,88	12,28	10,75	não
26/02/2019 18:15	345,60	2,40	118,51	0,33	44,33	15,93	10,75	sim
27/02/2019 05:55	357,50	1,59	131,65	0,21	218,31	12,89	8,67	não
27/02/2019 17:03	13,90	5,04	115,01	0,11	215,95	12,99	8,67	sim
28/02/2019 05:49	2,50	7,24	32,78	0,13	215,77	12,44	8,92	não
01/03/2019 16:52	29,90	7,37	118,13	0,29	92,72	12,86	15,42	não
02/03/2019 06:07	331,70	8,46	338,51	0,05	64,74	11,80	15,83	não
03/03/2019 05:55	346,10	4,84	116,42	0,09	62,38	6,40	16,83	não
03/03/2019 17:28	83,60	2,27	303,71	0,08	166,44	12,72	16,83	sim
05/03/2019 16:51	21,60	4,87	128,94	0,16	107,47	12,53	15,17	sim
06/03/2019 17:34	75,10	5,26	153,38	0,19	175,81	12,68	12,25	sim
09/03/2019 17:40	51,80	4,85	212,36	1,03	97,51	9,20	9,58	sim
11/03/2019 05:55	281,60	3,99	196,68	0,60	121,32	9,84	8,25	sim
11/03/2019 17:04	253,70	2,63	205,26	0,83	225,56	10,04	8,25	sim

Continuação da Tabela A 1

Data e hora	***DV* (°)**	***IV* (m/s)**	***DC* (°)**	***IC* (m/s)**	***DO* (°)**	***PP* (s)**	***TOG* (mg/L)**	***Feição***
12/03/2019 21:07	51,50	4,67	175,78	0,09	179,08	9,20	9,50	sim
16/03/2019 05:49	133,00	2,34	208,95	1,32	149,23	12,99	12,33	sim
19/03/2019 05:55	3,30	6,21	212,17	0,82	166,37	11,39	8,42	sim
01/04/2019 05:04	243,30	2,78	191,93	0,54	143,64	9,06	12,00	sim
09/04/2019 17:04	235,90	3,27	176,81	0,62	189,01	12,34	10,25	sim
10/04/2019 06:07	173,70	1,73	195,86	0,73	202,25	9,72	13,08	sim
15/04/2019 06:01	125,50	3,67	198,27	0,40	132,30	12,00	18,92	sim
15/04/2019 16:46	127,70	4,03	190,80	0,41	152,10	12,69	18,92	sim
15/04/2019 18:14	128,60	4,73	190,75	0,38	146,46	12,61	18,92	sim
16/04/2019 17:04	52,60	3,37	191,58	0,71	140,36	12,67	13,33	sim
20/04/2019 17:27	33,30	1,76	208,41	0,40	141,50	9,21	9,33	sim
21/04/2019 06:13	355,70	5,00	212,45	0,59	127,52	8,50	11,42	sim
21/04/2019 16:58	5,30	3,38	181,67	0,40	138,10	12,60	11,42	sim
23/04/2019 17:34	85,20	5,26	203,52	0,49	186,42	12,21	11,17	sim
26/04/2019 06:07	87,20	2,33	198,87	0,51	182,32	11,90	17,00	sim
05/05/2019 16:52	108,30	3,96	186,55	0,66	161,08	16,55	15,42	sim
10/05/2019 06:01	92,50	3,43	161,10	0,32	169,16	12,71	11,58	sim
14/05/2019 05:07	170,00	3,98	195,07	0,51	130,12	13,96	13,58	sim
14/05/2019 05:55	175,40	4,43	194,69	0,48	98,46	13,85	13,58	sim
18/05/2019 17:04	118,20	10,31	194,59	0,55	117,20	9,39	11,67	não
20/05/2019 05:20	114,10	3,35	187,38	0,43	114,53	11,41	15,42	não
20/05/2019 17:40	166,60	2,06	223,62	0,48	81,32	10,75	15,42	sim
21/05/2019 05:01	nan	nan	217,98	0,60	98,08	10,42	9,75	sim

Continuação da Tabela A 1

Data e hora	***DV* (°)**	***IV* (m/s)**	***DC* (°)**	***IC* (m/s)**	***DO* (°)**	***PP* (s)**	***TOG* (mg/L)**	***Feição***
22/05/2019 05:55	nan	nan	217,90	0,38	89,13	9,42	9,33	sim
23/05/2019 16:57	10,80	4,84	192,16	0,46	75,19	9,15	14,08	sim
01/06/2019 16:58	198,50	5,01	199,12	0,15	220,85	12,51	12,50	sim
02/06/2019 05:13	71,60	3,34	206,64	0,50	198,00	11,42	11,67	sim
08/06/2019 05:25	61,20	5,40	187,46	0,53	113,76	13,00	14,69	sim
20/06/2019 17:22	336,40	2,96	143,96	0,38	141,06	12,59	17,42	sim
23/06/2019 05:08	80,50	5,96	136,42	0,11	206,29	12,72	10,58	não
24/06/2019 05:05	335,00	4,79	175,65	0,21	143,76	12,85	12,58	não
24/06/2019 05:26	335,00	4,79	175,65	0,21	143,76	12,85	12,58	não
09/07/2019 17:04	110,90	4,18	299,56	0,09	nan	9,87	8,00	sim
10/07/2019 05:25	111,10	3,36	262,57	0,09	141,59	9,61	8,92	sim
11/07/2019 18:38	9,00	4,49	238,54	0,17	137,29	12,67	13,83	sim
12/07/2019 17:04	340,20	2,64	66,53	0,15	117,89	12,79	12,58	sim
14/07/2019 05:49	337,30	3,90	190,60	0,21	62,32	12,27	13,33	sim
14/07/2019 17:22	14,50	3,71	195,51	0,25	228,60	15,84	13,33	sim
27/07/2019 16:46	64,00	3,15	216,62	0,95	141,61	11,55	16,83	sim
28/07/2019 17:04	207,10	3,51	208,54	0,46	123,77	12,87	18,50	sim
29/07/2019 05:55	116,30	3,00	212,01	0,87	125,29	13,38	22,08	sim
28/08/2019 06:01	133,00	4,03	247,31	0,07	130,87	12,75	15,58	sim
05/09/2019 06:07	211,30	6,09	32,64	0,11	162,90	11,72	12,17	sim
19/10/2019 05:19	62,80	3,25	43,12	0,33	110,88	11,53	18,00	sim
22/10/2019 05:04	21,00	0,57	54,04	0,63	nan	nan	15,17	sim
22/10/2019 16:58	157,10	3,71	42,83	0,35	210,03	11,69	15,17	sim

Continuação da Tabela A 1

Data e hora	***DV* (°)**	***IV* (m/s)**	***DC* (°)**	***IC* (m/s)**	***DO* (°)**	***PP* (s)**	***TOG* (mg/L)**	***Feição***
05/01/2020 17:05	10,80	3,82	108,08	0,08	72,63	8,24	12,00	sim
15/01/2020 06:07	119,90	3,05	113,43	0,04	nan	nan	16,75	sim
21/01/2020 17:04	207,40	3,57	204,12	0,10	165,10	8,98	11,42	sim
11/02/2020 16:58	218,70	2,74	152,97	0,42	188,17	8,60	17,83	sim
11/02/2020 18:07	210,40	4,47	166,82	0,43	193,36	nan	17,83	sim
14/02/2020 05:49	319,30	3,23	162,74	0,05	169,77	nan	28,31	sim
15/02/2020 05:49	17,20	5,51	160,11	0,52	120,41	nan	20,31	sim
15/02/2020 16:52	52,60	6,63	161,23	0,69	100,91	nan	20,31	sim
26/02/2020 05:55	10,80	7,53	236,41	0,62	123,39	nan	12,17	sim
01/03/2020 05:49	73,20	3,30	205,91	1,13	166,58	12,70	13,67	sim
03/03/2020 18:06	103,40	2,09	114,28	0,63	153,82	9,10	13,92	não
05/03/2020 17:04	220,30	7,13	81,34	0,20	216,04	13,80	9,67	não
07/03/2020 17:38	184,60	4,04	278,81	0,07	nan	nan	10,25	sim
08/03/2020 16:46	130,10	3,47	nan	nan	174,72	11,40	8,00	sim
11/03/2020 05:19	133,90	3,91	292,71	0,31	109,81	9,80	7,67	sim
13/03/2020 05:55	94,70	4,52	53,51	0,41	117,67	12,50	19,08	sim
14/03/2020 05:04	76,40	4,22	62,05	0,90	165,40	11,40	18,75	sim
14/03/2020 16:58	93,00	4,30	44,18	0,65	149,02	6,10	18,75	sim
15/03/2020 16:52	45,60	6,26	220,30	1,41	92,47	12,80	14,67	sim
18/03/2020 05:49	256,50	2,94	251,48	0,95	197,25	11,70	14,17	sim
19/03/2020 05:43	0,30	4,19	252,89	0,75	279,63	12,30	17,42	sim
25/03/2020 05:55	183,70	3,67	246,59	0,42	155,05	14,30	11,00	sim
25/03/2020 17:04	125,20	2,68	310,69	0,81	178,57	12,80	11,00	sim

Continuação da Tabela A 1

Data e hora	***DV* (°)**	***IV* (m/s)**	***DC* (°)**	***IC* (m/s)**	***DO* (°)**	***PP* (s)**	***TOG* (mg/L)**	***Feição***
26/03/2020 05:04	122,90	3,73	324,76	0,30	178,27	12,70	9,75	sim
26/03/2020 05:49	133,90	3,69	280,70	0,09	172,76	12,80	9,75	sim
26/03/2020 17:22	136,10	3,33	287,22	nan	183,40	11,60	9,75	sim
27/03/2020 05:19	118,10	3,47	4,46	0,27	144,90	11,90	13,50	sim
29/03/2020 05:55	81,70	2,86	288,83	0,11	151,73	12,90	15,75	sim
30/03/2020 16:58	22,50	2,82	150,10	0,05	nan	nan	12,50	sim
31/03/2020 16:52	125,30	5,93	73,01	0,29	109,43	8,70	11,69	sim
01/04/2020 06:01	115,30	4,87	103,15	0,40	134,70	13,00	10,17	sim
02/04/2020 18:28	22,20	7,63	84,60	0,16	97,09	12,70	12,25	não
03/04/2020 16:52	333,90	3,31	270,51	0,57	32,31	7,60	11,42	sim
07/04/2020 05:05	34,10	7,78	214,02	0,32	168,17	12,20	10,67	não
11/04/2020 05:49	30,00	2,16	240,03	0,49	212,80	12,70	19,25	sim
14/04/2020 05:05	104,20	4,52	37,78	0,11	162,74	9,80	19,92	sim
14/04/2020 05:55	104,20	4,52	37,78	0,11	162,74	9,80	19,92	sim
15/04/2020 16:58	9,90	5,02	75,24	0,26	71,33	12,80	20,42	sim
16/04/2020 06:08	8,90	4,33	223,99	0,28	38,10	12,20	22,67	não
16/04/2020 18:10	162,30	7,24	147,40	0,59	202,82	11,10	22,67	não
18/04/2020 05:50	186,20	6,61	145,08	0,49	226,84	10,60	21,50	não
19/04/2020 05:04	188,00	4,68	141,58	0,59	217,72	10,70	18,58	sim
19/04/2020 05:49	176,10	4,57	166,28	0,64	226,15	10,50	18,58	sim
19/04/2020 16:52	186,40	3,89	150,69	0,43	200,94	10,60	18,58	sim
23/04/2020 05:25	172,10	3,74	71,48	0,40	43,78	9,20	15,58	sim
24/04/2020 06:01	178,60	5,26	66,62	0,48	172,81	12,00	16,75	sim

Continuação da Tabela A 1

Data e hora	***DV* (°)**	***IV* (m/s)**	***DC* (°)**	***IC* (m/s)**	***DO* (°)**	***PP* (s)**	***TOG* (mg/L)**	***Feição***
26/04/2020 17:04	108,70	2,79	139,51	0,10	119,31	12,60	16,83	sim
28/04/2020 16:52	132,80	3,97	200,73	0,17	173,65	9,20	13,17	sim
30/04/2020 05:55	41,90	4,79	227,25	0,41	135,02	13,00	13,83	sim
01/05/2020 16:58	227,50	2,86	239,10	0,82	210,39	14,50	13,00	sim
02/05/2020 16:52	4,40	4,93	118,48	0,45	225,12	11,70	15,17	sim
06/05/2020 16:52	6,00	6,88	123,90	0,50	48,87	11,50	9,67	sim
08/05/2020 05:08	178,50	7,64	64,00	0,24	229,42	12,40	14,25	não
09/05/2020 05:26	146,00	4,92	43,61	0,38	213,18	10,90	10,58	não
11/05/2020 06:02	57,50	6,27	85,28	0,15	116,03	11,40	14,83	não
17/05/2020 16:59	89,10	4,21	nan	nan	100,48	8,00	4,67	sim
18/05/2020 16:52	140,10	4,30	nan	nan	210,57	11,50	5,42	sim
18/05/2020 17:38	138,70	4,79	nan	nan	207,36	11,90	5,42	sim
19/05/2020 16:46	140,90	nan	nan	nan	nan	nan	5,92	sim
25/05/2020 05:25	17,90	1,80	nan	nan	83,27	9,20	13,27	sim
29/05/2020 17:23	182,90	3,10	nan	nan	204,10	11,60	8,67	sim
30/05/2020 16:52	19,90	4,92	nan	nan	168,23	11,50	9,75	sim
04/06/2020 17:34	125,90	6,62	nan	nan	173,97	11,50	10,58	sim
06/06/2020 05:04	8,70	3,50	nan	nan	141,88	15,80	8,69	sim
12/06/2020 17:04	5,30	5,99	nan	nan	43,43	7,50	5,92	sim
07/07/2020 17:04	20,10	4,78	142,20	0,10	116,67	9,80	4,42	sim
09/07/2020 16:52	224,00	4,25	56,61	0,17	31,07	6,40	4,75	não
13/07/2020 06:02	5,20	8,89	110,11	0,07	49,06	6,40	5,92	não
14/07/2020 06:02	0,50	7,62	52,26	0,16	62,98	11,20	8,42	não

Continuação da Tabela A 1

Data e hora	***DV* (°)**	***IV* (m/s)**	***DC* (°)**	***IC* (m/s)**	***DO* (°)**	***PP* (s)**	***TOG* (mg/L)**	***Feição***
14/07/2020 18:14	181,20	9,09	61,07	0,24	211,92	13,30	8,42	não
21/07/2020 16:52	19,00	6,07	157,98	0,16	80,68	9,00	5,67	não
23/07/2020 05:49	42,20	5,33	226,42	0,46	93,19	9,60	12,23	sim
23/07/2020 17:04	86,00	4,02	242,80	0,14	137,56	11,50	12,23	sim
24/07/2020 05:04	60,30	6,05	236,48	0,17	96,14	11,50	8,67	sim
24/07/2020 05:50	68,60	5,82	226,68	0,29	78,46	8,60	8,67	não
30/07/2020 06:02	296,60	1,91	158,75	0,37	152,64	8,00	8,50	não

Tabela A 2. Conjunto de dados usado para modelagem da extensão da feição oleosa

Data e hora	***DV* (°)**	***IV* (m/s)**	***DC* (°)**	***IC* (m/s)**	***DO* (°)**	***PP* (s)**	***TOG* (mg/L)**	***Feição* (MN)**
7/01/2018 18:15	8,8	3,02	257,59	0,22	77,79	9,02	17,58333	4,2
7/15/2018 16:58	10,9	4,98	203,91	0,12	0	7,57	10,75	7,5
7/18/2018 17:04	33,3	4,19	187,68	1	170,46	11,53	13,75	2,5
8/05/2018 05:19	241,6	2,31	199,17	0,19	125,26	8,12	17,76923	5,4
8/17/2018 16:58	132,6	3,43	191,17	0,49	206,97	12,26	6,5	1,8
9/09/2018 16:57	97,5	2,54	214,2	0,72	146,8	14,1	4,833333	6,2
9/11/2018 17:34	150	4	211,07	0,53	172,27	10,86	5,454545	8,9
9/11/2018 18:14	150	4	211,07	0,53	172,27	10,86	5,454545	8,7
9/17/2018 05:55	262,8	1,81	195,71	0,54	182,19	7,46	7,666667	5,7
9/30/2018 16:52	60,2	4,06	208,83	0,49	155,1	10,07	5,5	0,8
10/04/2018 16:58	242,9	2,39	203,01	0,92	113,18	8,7	11,91667	2,8
10/06/2018 05:55	163,6	3,87	202,92	0,71	190,77	9,2	4,166667	1,8
10/11/2018 17:46	354,7	2,65	211,51	1,09	75,86	8,43	10,69231	5,9

Continuação da Tabela A 2

Data e hora	***DV* (°)**	***IV* *(m/s)***	***DC* (°)**	***IC* *(m/s)***	***DO* (°)**	***PP* *(s)***	***TOG* *(mg/L)***	***Feição* *(MN)***
10/15/2018 05:49	179,9	2,72	171,18	0,62	207,57	12,13	11,08333	2
11/03/2018 05:55	75,5	3,36	209,24	0,6	150,69	11,47	10,66667	12,8
11/12/2018 17:57	93,4	6,56	223,11	0,29	209,95	8,14	14,30769	2,3
11/16/2018 18:04	214,4	3,22	222,49	0,29	75,71	6,28	17,5	2,7
11/20/2018 18:33	163,7	3,85	228,72	0,39	210,51	9	10,66667	1,8
11/26/2018 19:52	72,3	5,14	240,37	0,41	75,11	7,59	12,75	3,7
11/27/2018 06:55	255,2	2,4	237,19	0,23	185,17	8,01	11,91667	2,8
11/28/2018 06:49	106,6	1,93	239,19	0,28	166,01	13,41	8,666667	11
12/02/2018 06:04	272,6	3,94	232,92	0,32	39,31	8,05	11,16667	1,5
12/04/2018 07:01	145,3	4,32	205,73	0,27	205,65	9,35	9,75	2
12/11/2018 06:19	34,3	5	227,02	0,23	123,07	10,68	9,916667	1,7
12/14/2018 11:43	13,1	9,39	226,97	0,66	38,16	6,06	8,666667	1,6
12/25/2018 06:55	216,8	4,9	239,97	0,47	81,11	7,11	5	4
12/29/2018 06:55	332,6	6,23	236,07	0,39	34,17	6,83	10	2
12/30/2018 18:46	55,4	5,2	216,33	0,49	47,3	7,2	13,25	16,5
12/30/2018 19:07	55,4	5,2	216,33	0,49	47,3	7,2	13,25	7
12/31/2018 07:07	2,1	11,33	221,47	0,45	20,67	7,08	12,25	4,7
2/19/2019 18:18	8,8	5,45	204,67	0,14	198,19	12,81	4,75	1,4
2/20/2019 17:21	16,8	3,5	123,52	0,16	169,7	12,81	6,25	2,4
2/26/2019 18:15	345,6	2,4	118,51	0,33	44,33	15,93	10,75	4,5
2/27/2019 17:03	13,9	5,04	115,01	0,11	215,95	12,99	8,666667	3,3
3/03/2019 17:28	83,6	2,27	303,71	0,08	166,44	12,72	16,83333	1,5
3/05/2019 16:51	21,6	4,87	128,94	0,16	107,47	12,53	15,16667	1,3

Continuação da Tabela A 2

Data e hora	***DV (°)***	***IV (m/s)***	***DC (°)***	***IC (m/s)***	***DO (°)***	***PP (s)***	***TOG (mg/L)***	***Feição (MN)***
3/06/2019 17:34	75,1	5,26	153,38	0,19	175,81	12,68	12,25	1,5
3/09/2019 17:40	51,8	4,85	212,36	1,03	97,51	9,2	9,583333	3,3
3/11/2019 05:55	281,6	3,99	196,68	0,6	121,32	9,84	8,25	10,7
3/11/2019 17:04	253,7	2,63	205,26	0,83	225,56	10,04	8,25	9
3/12/2019 21:07	51,5	4,67	175,78	0,09	179,08	9,2	9,5	2,2
3/16/2019 05:49	133	2,34	208,95	1,32	149,23	12,99	12,33333	14
3/19/2019 05:55	3,3	6,21	212,17	0,82	166,37	11,39	8,416667	11,7
4/01/2019 05:04	243,3	2,78	191,93	0,54	143,64	9,06	12	2,8
4/09/2019 17:04	235,9	3,27	176,81	0,62	189,01	12,34	10,25	4,4
4/10/2019 06:07	173,7	1,73	195,86	0,73	202,25	9,72	13,07692	1,1
4/15/2019 06:01	125,5	3,67	198,27	0,4	132,3	12	18,91667	5,3
4/15/2019 16:46	127,7	4,03	190,8	0,41	152,1	12,69	18,91667	10,7
4/15/2019 18:14	128,6	4,73	190,75	0,38	146,46	12,61	18,91667	6,6
4/16/2019 17:04	52,6	3,37	191,58	0,71	140,36	12,67	13,33333	2
4/20/2019 17:27	33,3	1,76	208,41	0,4	141,5	9,21	9,333333	4,9
4/21/2019 06:13	355,7	5	212,45	0,59	127,52	8,5	11,41667	2,5
4/21/2019 16:58	5,3	3,38	181,67	0,4	138,1	12,6	11,41667	3,4
4/23/2019 17:34	85,2	5,26	203,52	0,49	186,42	12,21	11,16667	2
4/26/2019 06:07	87,2	2,33	198,87	0,51	182,32	11,9	17	7,4
5/05/2019 16:52	108,3	3,96	186,55	0,66	161,08	16,55	15,41667	5,8
5/10/2019 06:01	92,5	3,43	161,1	0,32	169,16	12,71	11,58333	2
5/14/2019 05:07	170	3,98	195,07	0,51	130,12	13,96	13,58333	2,3
5/14/2019 05:55	175,4	4,43	194,69	0,48	98,46	13,85	13,58333	2

Continuação da Tabela A 2

Data e hora	***DV* (°)**	***IV* *(m/s)***	***DC* (°)**	***IC* *(m/s)***	***DO* (°)**	***PP* *(s)***	***TOG* *(mg/L)***	***Feição* *(MN)***
5/20/2019 17:40	166,6	2,06	223,62	0,48	81,32	10,75	15,41667	3,5
5/21/2019 05:01	nan	nan	217,98	0,6	98,08	10,42	9,75	6
5/22/2019 05:55	nan	nan	217,9	0,38	89,13	9,42	9,333333	20
5/23/2019 16:57	10,8	4,84	192,16	0,46	75,19	9,15	14,07692	2,5
6/01/2019 16:58	198,5	5,01	199,12	0,15	220,85	12,51	12,5	3,4
6/02/2019 05:13	71,6	3,34	206,64	0,5	198	11,42	11,66667	3
6/08/2019 05:25	61,2	5,4	187,46	0,53	113,76	13	14,69231	1,2
6/20/2019 17:22	336,4	2,96	143,96	0,38	141,06	12,59	17,41667	2
7/10/2019 05:25	111,1	3,36	262,57	0,09	141,59	9,61	8,916667	6,3
7/11/2019 18:38	9	4,49	238,54	0,17	137,29	12,67	13,83333	1,6
7/12/2019 17:04	340,2	2,64	66,53	0,15	117,89	12,79	12,58333	1,7
7/14/2019 05:49	337,3	3,9	190,6	0,21	62,32	12,27	13,33333	2,5
7/14/2019 17:22	14,5	3,71	195,51	0,25	228,6	15,84	13,33333	4,4
7/27/2019 16:46	64	3,15	216,62	0,95	141,61	11,55	16,83333	8,7
7/28/2019 17:04	207,1	3,51	208,54	0,46	123,77	12,87	18,5	1,3
7/29/2019 05:55	116,3	3	212,01	0,87	125,29	13,38	22,07692	2,7
8/28/2019 06:01	133	4,03	247,31	0,07	130,87	12,75	15,58333	6,4
9/05/2019 06:07	211,3	6,09	32,64	0,11	162,9	11,72	12,16667	1,2
10/19/2019 05:19	62,8	3,25	43,12	0,33	110,88	11,53	18	4,9
10/22/2019 16:58	157,1	3,71	42,83	0,35	210,03	11,69	15,16667	1,6
1/05/2020 17:05	10,8	3,82	108,08	0,08	72,63	8,24	12	3
1/21/2020 17:04	207,4	3,57	204,12	0,1	165,1	8,98	11,41667	1,27
2/11/2020 16:58	218,7	2,74	152,97	0,42	188,17	8,6	17,83333	1,2

Continuação da Tabela A 2

Data e hora	***DV* (°)**	***IV* (m/s)**	***DC* (°)**	***IC* (m/s)**	***DO* (°)**	***PP* (s)**	***TOG* (mg/L)**	***Feição* (MN)**
2/11/2020 18:07	210,4	4,47	166,82	0,43	193,36	nan	17,83333	1,1
2/14/2020 05:49	319,3	3,23	162,74	0,05	169,77	nan	28,30769	3,7
2/15/2020 05:49	17,2	5,51	160,11	0,52	120,41	nan	20,30769	2,2
2/15/2020 16:52	52,6	6,63	161,23	0,69	100,91	nan	20,30769	2,7
2/26/2020 05:55	10,8	7,53	236,41	0,62	123,39	nan	12,16667	3,6
3/01/2020 05:49	73,2	3,3	205,91	1,13	166,58	12,7	13,66667	1,2
3/11/2020 05:19	133,9	3,91	292,71	0,31	109,81	9,8	7,666667	5,7
3/13/2020 05:55	94,7	4,52	53,51	0,41	117,67	12,5	19,08333	4,2
3/14/2020 05:04	76,4	4,22	62,05	0,9	165,4	11,4	18,75	8,5
3/14/2020 16:58	93	4,3	44,18	0,65	149,02	6,1	18,75	25,7
3/15/2020 16:52	45,6	6,26	220,3	1,41	92,47	12,8	14,66667	1,18
3/18/2020 05:49	256,5	2,94	251,48	0,95	197,25	11,7	14,16667	13,2
3/19/2020 05:43	0,3	4,19	252,89	0,75	279,63	12,3	17,41667	2,9
3/25/2020 05:55	183,7	3,67	246,59	0,42	155,05	14,3	11	4,6
3/25/2020 17:04	125,2	2,68	310,69	0,81	178,57	12,8	11	4
3/26/2020 05:04	122,9	3,73	324,76	0,3	178,27	12,7	9,75	8,8
3/26/2020 17:22	136,1	3,33	287,22	nan	183,4	11,6	9,75	11,09
3/27/2020 05:19	118,1	3,47	4,46	0,27	144,9	11,9	13,5	7,2
3/29/2020 05:55	81,7	2,86	288,83	0,11	151,73	12,9	15,75	6,6
3/31/2020 16:52	125,3	5,93	73,01	0,29	109,43	8,7	11,69231	3,4
4/01/2020 06:01	115,3	4,87	103,15	0,4	134,7	13	10,16667	8
4/03/2020 16:52	333,9	3,31	270,51	0,57	32,31	7,6	11,41667	1,9
4/11/2020 05:49	30	2,16	240,03	0,49	212,8	12,7	19,25	6,8

Continuação da Tabela A 2

Data e hora	***DV* (°)**	***IV* *(m/s)***	***DC* (°)**	***IC* *(m/s)***	***DO* (°)**	***PP* *(s)***	***TOG* *(mg/L)***	***Feição* *(MN)***
4/14/2020 05:05	104,2	4,52	37,78	0,11	162,74	9,8	19,91667	3,7
4/14/2020 05:55	104,2	4,52	37,78	0,11	162,74	9,8	19,91667	5
4/15/2020 16:58	9,9	5,02	75,24	0,26	71,33	12,8	20,41667	1,9
4/19/2020 05:04	188	4,68	141,58	0,59	217,72	10,7	18,58333	5,9
4/19/2020 05:49	176,1	4,57	166,28	0,64	226,15	10,5	18,58333	5
4/19/2020 16:52	186,4	3,89	150,69	0,43	200,94	10,6	18,58333	1,9
4/23/2020 05:25	172,1	3,74	71,48	0,4	43,78	9,2	15,58333	6,8
4/24/2020 06:01	178,6	5,26	66,62	0,48	172,81	12	16,75	1,2
4/26/2020 17:04	108,7	2,79	139,51	0,1	119,31	12,6	16,83333	8,7
4/28/2020 16:52	132,8	3,97	200,73	0,17	173,65	9,2	13,16667	1,9
4/30/2020 05:55	41,9	4,79	227,25	0,41	135,02	13	13,83333	5,6
5/01/2020 16:58	227,5	2,86	239,1	0,82	210,39	14,5	13	2,5
5/02/2020 16:52	4,4	4,93	118,48	0,45	225,12	11,7	15,16667	1,4
5/06/2020 16:52	6	6,88	123,9	0,5	48,87	11,5	9,666667	1,45
7/07/2020 17:04	20,1	4,78	142,2	0,1	116,67	9,8	4,416667	1,29
7/23/2020 05:49	42,2	5,33	226,42	0,46	93,19	9,6	12,23077	3,3
7/23/2020 17:04	86	4,02	242,8	0,14	137,56	11,5	12,23077	7
7/24/2020 05:04	60,3	6,05	236,48	0,17	96,14	11,5	8,666667	1,3

ANEXO B

Artigo: "*Optimization of the Resistance Spot Welding Process of 22MnB5-Galvannealed Steel Using Response Surface Methodology and Global Criterion Method Based on Principal Components Analysis*". Publicado na "***Metals***".

metals

MDPI

Article

Optimization of the Resistance Spot Welding Process of 22MnB5-Galvannealed Steel Using Response Surface Methodology and Global Criterion Method Based on Principal Components Analysis

Robson Ribeiro *, Estevão Luiz Romão, Eduardo Luz, José Henrique Gomes and Sebastião Costa

Institute of Industrial Engineering and Management, Federal University of Itajubá, Itajubá 37500-903, Brazil; estevaoromao@gmail.com (E.L.R.); eduardoluz@mescaldeiraria.com.br (E.L.); ze_henriquefg@unifei.edu.br (J.H.G.); sccosta@unifei.edu.br (S.C.)
* Correspondence: robsonricardosoribeiro@gmail.com

Received: 20 July 2020; Accepted: 2 September 2020; Published: 7 October 2020

check for updates

Abstract: The 22MnB5-galvannealed steel is extensively used in the hot stamping process to produce car anti-collision structure parts. Furthermore, the resistance spot welding (RSW) is an important process in the automobile industry, especially in body construction, and the 22MnB5-galvannealed steels are a big challenge for the joining methods because their microstructure and mechanical properties are different from those of the conventional steels. In view of this, the present paper aims to optimize the parameters of the RSW process of the 22MnB5-galvannealed steel. Initially, the goal was to remove the galvannealed coating and in the next stage, the following responses were maximized: the nugget width, the nugget cross-sectional area, the penetration, the strength, the joint efficiency, and the energy absorption, whereas the indentation, heat affected zone and separation were used as constraints. The process parameters selected were the effective welding time, the effective welding current, the quenching time, and the upslope time. Response surface methodology (RSM) was applied jointly with the global criterion method based on principal components. The results of the multiobjective optimization are close to the individual targets for each response, highlighting the importance of considering the correlation structure presented in the responses.

Keywords: resistance spot welding; 22MnB5-galvannealed; multiobjective optimization

1. Introduction

The global automobile industry has faced many challenges in different areas, such as energy, gas emission, security and accessibility. The reduction of the vehicle mass is one of the main strategies used to overcome these challenges. However, to maximize the reduction of the vehicle mass, materials with metallurgic properties, which enables the combination of resistance and lightness, should replace the conventionally used low-carbon steels [1].

Among the large number of materials developed for this purpose, the advanced high strength steels (AHSS) have become a promising alternative to reduce the weight without affecting the structure of the vehicle [2]. According to [3], the 22MnB5-galvannealed steel stands out among the other AHSS. It is largely used in hot stamping process because of its good aptitude for quenching, attaining a shear strength resistance around 1500 MPa [4–6]. Besides, it also has a superficial layer consisting of iron and zinc (Fe-Zn), resistant to the oxidation, which protects the structural components when exposed to the environment [3].

Artigo: "*A new multiobjective optimization with elliptical constraints approach for nonlinear models implemented in a stainless steel cladding process*". Publicado na "***International Journal of Advanced Manufacturing Technology***".

The International Journal of Advanced Manufacturing Technology (2021) 113:1469–1484
https://doi.org/10.1007/s00170-020-06581-3

ORIGINAL ARTICLE

A new multiobjective optimization with elliptical constraints approach for nonlinear models implemented in a stainless steel cladding process

Eduardo Rivelino Luz[1] · Estevão Luiz Romão[1] · Simone Carneiro Streitenberger[1] · José Henrique Freitas Gomes[1] · Anderson Paulo de Paiva[1] · Pedro Paulo Balestrassi[1]

Received: 28 September 2020 / Accepted: 28 December 2020 / Published online: 6 February 2021

Abstract
This paper proposes a new multiobjective optimization with elliptical constraints approach for nonlinear models implemented in a cladding process of ABNT 1020 carbon steel plate using austenitic ABNT 316L stainless steel cored wire. Stainless steel stands out among the cladding materials as it allows obtaining surfaces with determined desirable characteristics from lower cost materials. This kind of process may be controlled by a relatively small number of input variables, i.e., the wire feed rate (*WF*), voltage (*V*), welding speed (*WS*), and the distance from the contact tip to the workpiece (*N*). Besides that, many outputs can be evaluated and simultaneously optimized. In the present paper, dilution (*D*), yield (*Y*), convexity index (*CI*), and penetration index (*PI*) were investigated. In order to consider the problem's multivariate nature, techniques such as factor analysis and Bonferroni's multivariate intervals were applied combined with elliptical constraints. The response variables were mathematically modeled using Poisson regression, and the obtained results were satisfactory since accurate models were achieved. The normal boundary intersection (NBI) method produced a set of viable configurations for the input variables that allows the experimenter to encounter the best system setup regarding the importance level of each response. Feasible and non-dominated solutions were found which means that the best possible scenario considering all the constraints was reached.

Keywords Cladding · Design of experiments · Multiobjective constrained optimization · Factor analysis · Simultaneous confidence intervals

1 Introduction

The use of welding as a manufacturing or maintenance process in various segments of the industry has been a milestone for its growth and strengthening. Industries in general have been constantly looking for alternatives in order to reduce costs by minimizing wear and tear in their equipment [1]. For instance, stainless steels are generally deposited on surfaces of carbon steels or low alloy steels, producing a layer with anticorrosive and resistant properties that are necessary to withstand environments subject to high corrosion. This is one of the applications of cladding [2].

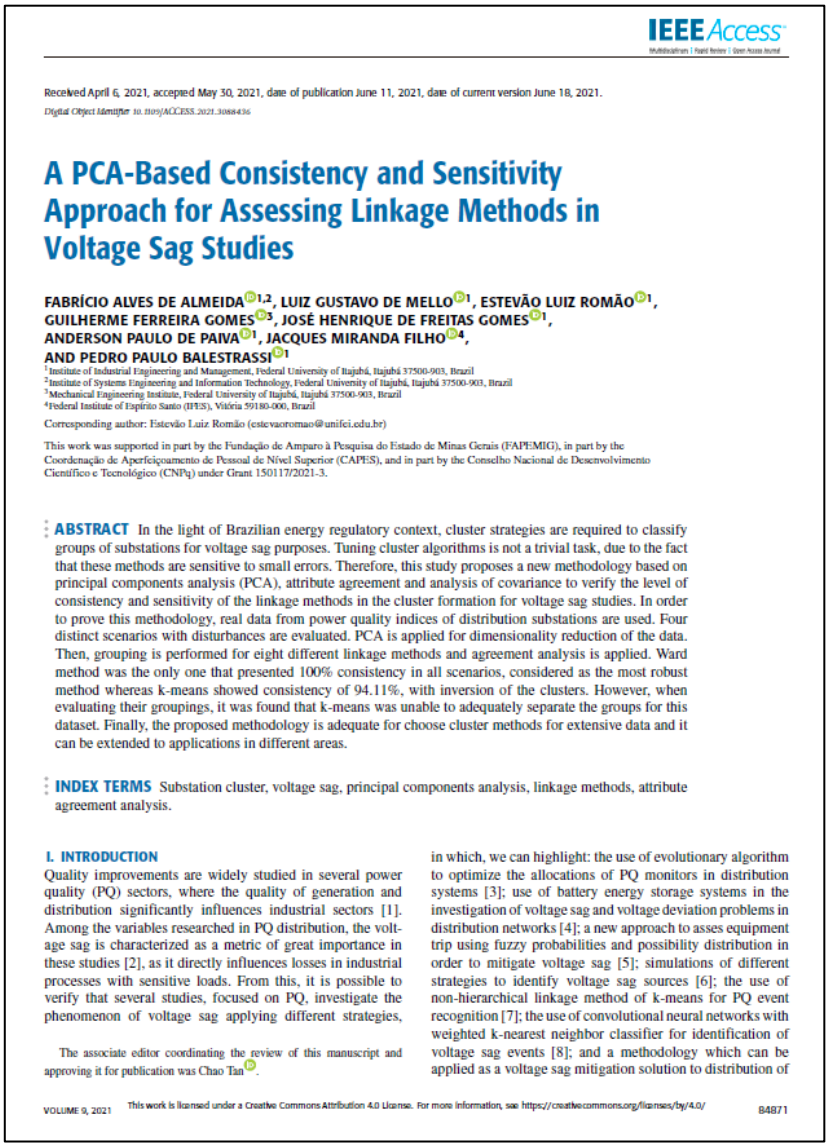
✉ Eduardo Rivelino Luz
eluz7777@gmail.com

[1] Institute of Industrial Engineering and Management, Federal University of Itajubá, Itajubá, MG, Brazil

Cladding process is defined as the deposition of a sufficiently thick layer of some weld metal of interest on a carbon steel or low alloy surface to make it resistant to corrosion or wear [3]. It is generally applied to extend the useful life of parts that do not have all the necessary properties for a specific application and to recover elements or materials that no longer have certain characteristics required by the process or are in a state of wear or corrosion [3–6].

The results of this process have made this application quite attractive, insofar as surfaces that are resistant to corrosive environments can be produced from common materials at a lower cost compared to the use of purely stainless steel components. It guarantees the reuse of the original material that would go for disposal, and it may result in a manufacturing cost reduction, besides making the process more sustainable [7–9]. This technique extends among the most diverse types of industries, as oil, chemical, food, agricultural, nuclear, naval, railway, and civil construction [6, 10].

Artigo: "*A PCA-Based Consistency and Sensitivity Approach for Assessing Linkage Methods in Voltage Sag Studies*". Publicado na "***IEEE Access***".



IEEE Access

Received April 6, 2021, accepted May 30, 2021, date of publication June 11, 2021, date of current version June 18, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3088436

A PCA-Based Consistency and Sensitivity Approach for Assessing Linkage Methods in Voltage Sag Studies

FABRÍCIO ALVES DE ALMEIDA[1,2], LUIZ GUSTAVO DE MELLO[1], ESTEVÃO LUIZ ROMÃO[1], GUILHERME FERREIRA GOMES[3], JOSÉ HENRIQUE DE FREITAS GOMES[1], ANDERSON PAULO DE PAIVA[1], JACQUES MIRANDA FILHO[4], AND PEDRO PAULO BALESTRASSI[1]

[1]Institute of Industrial Engineering and Management, Federal University of Itajubá, Itajubá 37500-903, Brazil
[2]Institute of Systems Engineering and Information Technology, Federal University of Itajubá, Itajubá 37500-903, Brazil
[3]Mechanical Engineering Institute, Federal University of Itajubá, Itajubá 37500-903, Brazil
[4]Federal Institute of Espírito Santo (IFES), Vitória 59180-000, Brazil

Corresponding author: Estevão Luiz Romão (estevaoromao@unifei.edu.br)

This work was supported in part by the Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG), in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), and in part by the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) under Grant 150117/2021-3.

ABSTRACT In the light of Brazilian energy regulatory context, cluster strategies are required to classify groups of substations for voltage sag purposes. Tuning cluster algorithms is not a trivial task, due to the fact that these methods are sensitive to small errors. Therefore, this study proposes a new methodology based on principal components analysis (PCA), attribute agreement and analysis of covariance to verify the level of consistency and sensitivity of the linkage methods in the cluster formation for voltage sag studies. In order to prove this methodology, real data from power quality indices of distribution substations are used. Four distinct scenarios with disturbances are evaluated. PCA is applied for dimensionality reduction of the data. Then, grouping is performed for eight different linkage methods and agreement analysis is applied. Ward method was the only one that presented 100% consistency in all scenarios, considered as the most robust method whereas k-means showed consistency of 94.11%, with inversion of the clusters. However, when evaluating their groupings, it was found that k-means was unable to adequately separate the groups for this dataset. Finally, the proposed methodology is adequate for choose cluster methods for extensive data and it can be extended to applications in different areas.

INDEX TERMS Substation cluster, voltage sag, principal components analysis, linkage methods, attribute agreement analysis.

I. INTRODUCTION

Quality improvements are widely studied in several power quality (PQ) sectors, where the quality of generation and distribution significantly influences industrial sectors [1]. Among the variables researched in PQ distribution, the voltage sag is characterized as a metric of great importance in these studies [2], as it directly influences losses in industrial processes with sensitive loads. From this, it is possible to verify that several studies, focused on PQ, investigate the phenomenon of voltage sag applying different strategies, in which, we can highlight: the use of evolutionary algorithm to optimize the allocations of PQ monitors in distribution systems [3]; use of battery energy storage systems in the investigation of voltage sag and voltage deviation problems in distribution networks [4]; a new approach to asses equipment trip using fuzzy probabilities and possibility distribution in order to mitigate voltage sag [5]; simulations of different strategies to identify voltage sag sources [6]; the use of non-hierarchical linkage method of k-means for PQ event recognition [7]; the use of convolutional neural networks with weighted k-nearest neighbor classifier for identification of voltage sag events [8]; and a methodology which can be applied as a voltage sag mitigation solution to distribution of

The associate editor coordinating the review of this manuscript and approving it for publication was Chao Tan.

Artigo: "*A Multiobjective optimization of the welding process in aluminum alloy (AA) 6063 T4 tubes used in corona rings through normal boundary intersection and multivariate techniques*". Publicado no "***The International Journal of Advanced Manufacturing Technology***".

The International Journal of Advanced Manufacturing Technology
https://doi.org/10.1007/s00170-021-07761-5

ORIGINAL ARTICLE

A multiobjective optimization of the welding process in aluminum alloy (AA) 6063 T4 tubes used in corona rings through normal boundary intersection and multivariate techniques

Eduardo Rivelino Luz[1] · Estevão Luiz Romão[1] · Simone Carneiro Streitenberger[1] · Leonardo Ribeiro Mandilha[1] · Anderson Paulo de Paiva[1] · Pedro Paulo Balestrassi[1]

Received: 15 May 2021 / Accepted: 20 July 2021

Abstract
The welding process in aluminum is a complex process that commonly presents several issues such as weld bead discontinuity, cracks, and lack of penetration. Thus, an accurate specification of the parameters in order to achieve optimal values for the investigated responses is aimed by the industry. The present paper proposes the application of a multiobjective optimization approach considering multivariate constraints based on the simultaneous confidence intervals and the elliptical region of the correlated data. Structured experiments for the welding process of aluminum alloy (AA) 6063 TA tubes used in corona rings were performed according to a face-centered composite design with 4 factors, wire feed rate (Wf), arc voltage (V), contact tip to the workpiece distance (Ct), and motor frequency (Fr), resulting in 31 experiments. Poisson regression was applied to model the values of yield (Y), dilution (D), reinforcement index (RI), and penetration index (PI), allowing to estimate the optimal individual values with regard to the multivariate constraints. Rotated factor scores were obtained in order to replace the original data, and therefore, the factor multivariate square error was used as objective functions to be minimized through normal boundary intersection method. A satisfactory weld bead with large values of PI, D, and Y and a small value of RI was reached as prespecified by the manager of the process.

Keywords Aluminum welding · Design of experiments · Multiobjective constrained optimization · Factor analysis · Simultaneous confidence intervals

1 Introduction

Corona rings are used to improve the performance of the insulator strings. Besides reducing corona discharges, the associated audible noise, and radio and television interference levels, they also improve the voltage distribution along the insulator string by reducing the percentage of the voltage on the nearest unit to the power transmission line. Moreover, corona rings also alleviate corona degradation of nonceramic materials. Structurally, they are toroidal shaped metallic rings which are fixed at the end of bushings and insulator strings. Also called anti-corona rings, they can even be used to prevent corona discharge that occurs in high voltage power lines. This discharge, or corona loss, is a significant issue in very high voltage power lines, causing power loss [1, 2].

In order to increase the lifetime of a composite insulator, mainly when it carries high voltages like over 230 kV, a field stress control ring, such as the corona ring, is necessary. Furthermore, it is important that a great coherence between the insulator and the corona ring in terms of the design and construction exists; i.e., the dimensions of the corona ring should be well specified during the insulator construction [3].

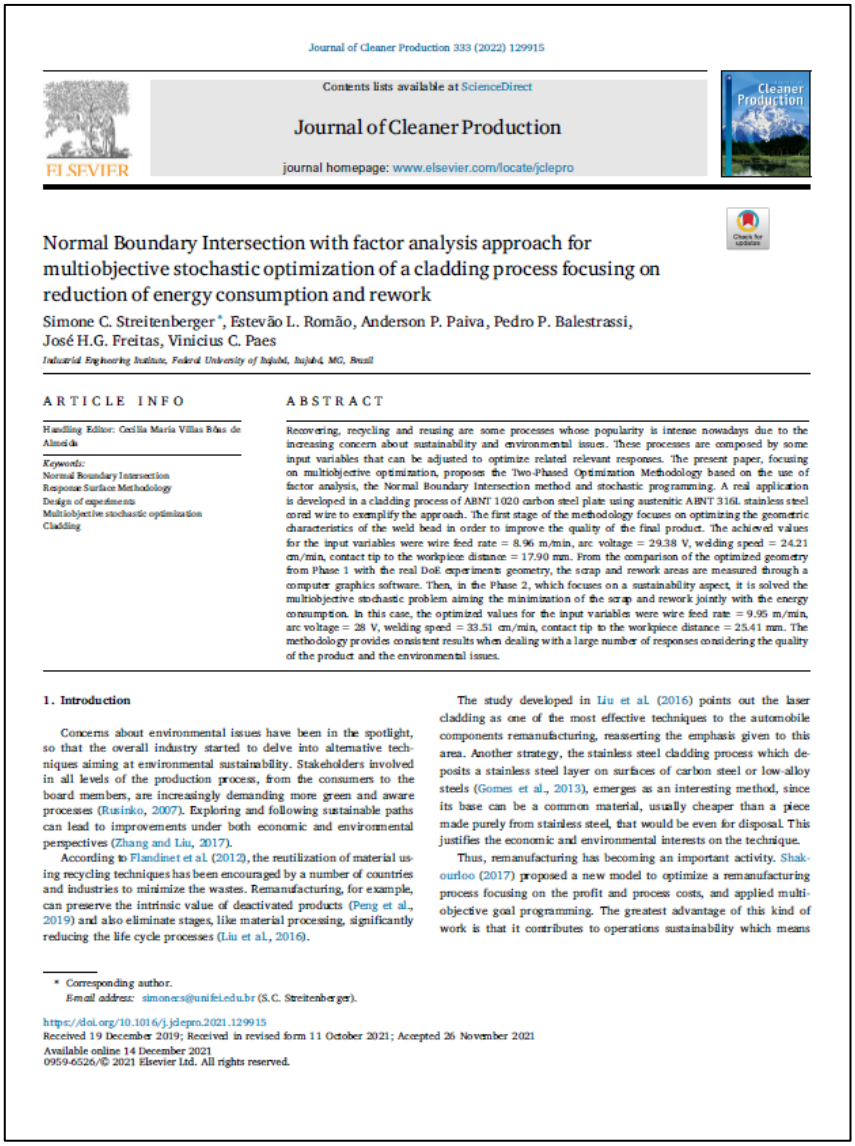
Constraints and better operating conditions are applied for better efficiency and organizational effectiveness in the management of the gas metal arc welding (GMAW) process, better known as metal inert gas (MIG) welding, of anti-corona protection rings. Considering a project from a Brazilian company, these rings are manufactured using aluminum alloy (AA) 6063 T4. Specifically for this study, a tube with a diameter of 100 mm and thickness of 2 mm was chosen. The

✉ Eduardo Rivelino Luz
eluz7777@gmail.com

[1] Institute of Industrial Engineering and Management, Federal University of Itajubá, Itajubá, MG, Brazil

Published online: 09 August 2021

Artigo: "*Normal Boundary Intersection with factor analysis approach for multiobjective stochastic optimization of a cladding process focusing on reduction of energy consumption and rework*". Publicado no "***Journal of Cleaner Production***".



Journal of Cleaner Production 333 (2022) 129915

Contents lists available at ScienceDirect

Journal of Cleaner Production

journal homepage: www.elsevier.com/locate/jclepro

Normal Boundary Intersection with factor analysis approach for multiobjective stochastic optimization of a cladding process focusing on reduction of energy consumption and rework

Simone C. Streitenberger *, Estevão L. Romão, Anderson P. Paiva, Pedro P. Balestrassi, José H.G. Freitas, Vinicius C. Paes

Industrial Engineering Institute, Federal University of Itajubá, Itajubá, MG, Brazil

ARTICLE INFO

Handling Editor: Cecília Maria Villas Bôas de Almeida

Keywords:
Normal Boundary Intersection
Response Surface Methodology
Design of experiments
Multiobjective stochastic optimization
Cladding

ABSTRACT

Recovering, recycling and reusing are some processes whose popularity is intense nowadays due to the increasing concern about sustainability and environmental issues. These processes are composed by some input variables that can be adjusted to optimize related relevant responses. The present paper, focusing on multiobjective optimization, proposes the Two-Phased Optimization Methodology based on the use of factor analysis, the Normal Boundary Intersection method and stochastic programming. A real application is developed in a cladding process of ABNT 1020 carbon steel plate using austenitic ABNT 316L stainless steel cored wire to exemplify the approach. The first stage of the methodology focuses on optimizing the geometric characteristics of the weld bead in order to improve the quality of the final product. The achieved values for the input variables were wire feed rate = 8.96 m/min, arc voltage = 29.38 V, welding speed = 24.21 cm/min, contact tip to the workpiece distance = 17.90 mm. From the comparison of the optimized geometry from Phase 1 with the real DoE experiments geometry, the scrap and rework areas are measured through a computer graphics software. Then, in the Phase 2, which focuses on a sustainability aspect, it is solved the multiobjective stochastic problem aiming the minimization of the scrap and rework jointly with the energy consumption. In this case, the optimized values for the input variables were wire feed rate = 9.95 m/min, arc voltage = 28 V, welding speed = 33.51 cm/min, contact tip to the workpiece distance = 25.41 mm. The methodology provides consistent results when dealing with a large number of responses considering the quality of the product and the environmental issues.

1. Introduction

Concerns about environmental issues have been in the spotlight, so that the overall industry started to delve into alternative techniques aiming at environmental sustainability. Stakeholders involved in all levels of the production process, from the consumers to the board members, are increasingly demanding more green and aware processes (Rusinko, 2007). Exploring and following sustainable paths can lead to improvements under both economic and environmental perspectives (Zhang and Liu, 2017).

According to Flandinet et al. (2012), the reutilization of material using recycling techniques has been encouraged by a number of countries and industries to minimize the wastes. Remanufacturing, for example, can preserve the intrinsic value of deactivated products (Peng et al., 2019) and also eliminate stages, like material processing, significantly reducing the life cycle processes (Liu et al., 2016).

The study developed in Liu et al. (2016) points out the laser cladding as one of the most effective techniques to the automobile components remanufacturing, reasserting the emphasis given to this area. Another strategy, the stainless steel cladding process which deposits a stainless steel layer on surfaces of carbon steel or low-alloy steels (Gomes et al., 2013), emerges as an interesting method, since its base can be a common material, usually cheaper than a piece made purely from stainless steel, that would be even for disposal. This justifies the economic and environmental interests on the technique.

Thus, remanufacturing has becoming an important activity. Shakourloo (2017) proposed a new model to optimize a remanufacturing process focusing on the profit and process costs, and applied multiobjective goal programming. The greatest advantage of this kind of work is that it contributes to operations sustainability which means

* Corresponding author.
Email address: simoners@unifei.edu.br (S.C. Streitenberger).

https://doi.org/10.1016/j.jclepro.2021.129915
Received 19 December 2019; Received in revised form 11 October 2021; Accepted 26 November 2021
Available online 14 December 2021
0959-6526/

Artigo: "*Combining machine learning techniques with Kappa-Kendall indexes for robust hard-cluster assessment in substation pattern recognition*". Publicado na "***Electric Power Systems Research***"

Electric Power Systems Research 206 (2022) 107778

Contents lists available at ScienceDirect

Electric Power Systems Research

journal homepage: www.elsevier.com/locate/epsr

Combining machine learning techniques with Kappa–Kendall indexes for robust hard-cluster assessment in substation pattern recognition

Fabricio Alves de Almeida [a,*], Estevão Luiz Romão [b], Guilherme Ferreira Gomes [c], José Henrique de Freitas Gomes [b], Anderson Paulo de Paiva [b], Jacques Miranda Filho [d], Pedro Paulo Balestrassi [a,b]

[a] Institute of Electrical Systems and Energy, Federal University of Itajubá, Brazil
[b] Institute of Industrial Engineering and Management, Federal University of Itajubá, Brazil
[c] Mechanical Engineering Institute, Federal University of Itajubá, Brazil
[d] IFES Federal Institute of Espírito Santo, Vitória, Brazil

ARTICLE INFO

Keywords:
Machine learning
Kappa-Kendall
Rotated factor analysis
Voltage sag
Power quality substation
Pattern recognition

ABSTRACT

This study proposes a method that combines different machine learning and lean six sigma techniques to calibrate cluster analysis through linkage methods. The power quality indexes of substations in Brazil, which are of interest to regulatory agencies, are used. The method uses the random forest mixed with rotated factor analysis to filter, minimize, and improve the interpretation of latent information. Variability scenarios are created using the Monte Carlo simulation to assess the stability of the cluster analysis using the design of experiments and the Kappa–Kendall indexes. The Ward method shows a better consistency in all scenarios and a better discriminatory power among the clusters. The optimal result is used to predict different scenarios with high levels of variability (5, 10, and 15%) by comparing the behaviors of different supervised machine learning techniques for classification. The results show that the k-nearest neighbors, support vector classifier, and logistic regression approaches can accurately predict, even in scenarios with high variability in the dataset.

1. Introduction

Advanced statistical techniques have been widely investigated in power quality (PQ) studies [1], promoting the development of new technologies and supporting decision making [2]. The advent of computation has provided the creation and improvement of mathematical/machine learning approaches in strategic sectors, such as energy generation and distribution, thus, impacting the industrial sector significantly [1]. Voltage sag is a significant characteristic of PQ distribution [3] and is a variable caused by the short-duration voltage variation. Additionally, this variable economically impacts the production processes because industrial processes have sensitive loads. Several existing studies investigated the voltage sag phenomenon [4–8].

As an object of study, regulatory agencies consider voltage sag and other characteristics to classify PQ substations based on the number of voltage sag events. Some studies have used exploratory techniques combined with cluster analysis (CA) to group substations based on the PQ [1,9]. These studies were based on the regulatory agencies' need to assess and control the PQ. In these studies, specific techniques were applied in view of the multivariate characteristics of the data.

Multivariate techniques were used to analyse a dataset with multiple correlated characteristics [10]. Factor analysis (FA), an exploratory strategy, is one the most robust techniques. This technique transforms several variables into a few common factors, thereby reducing the data dimensionality. This technique also allows rotation of factor loads, simplifying the load matrix and creating an easy-to-interpret structure [11]. Another multivariate strategy widely used in the electricity sector is the CA. This technique creates clusters based on the level of similarity using techniques such as the hierarchical and non-hierarchical linkage methods. Both approaches can recognise patterns of observations based on the available characteristics. These strategies were applied in several studies on PQ [12–16], highlighting their importance.

The most used and significant linkage methods are: k-means, Ward, single, average, complete, median, centroid, and McQuitty. However, many authors arbitrarily employ linkage methods obtaining unsatisfactory results because linkage methods, which are sensitive to outliers,

* Corresponding author.
E-mail address: fabricio-almeida@unifei.edu.br (F.A. Almeida).

https://doi.org/10.1016/j.epsr.2022.107778
Received 29 August 2020; Received in revised form 23 November 2021; Accepted 4 January 2022