VETOR

Uso de Métodos de Aprendizado de Máquina e Algoritmo Genético para Predição de TOC e Classificação de Litologia☆

Use of Machine Learning Methods and Genetic Algorithm for TOC Prediction and Lithology Classification

Juliana da Costa Cabral[1], Clovis Antonio da Silva[1], Grazione de Souza[2], Camila Martins Saporetti[2,†]

[1]*Pós-Graduação em Modelagem Computacional, Universidade do Estado do Rio de Janeiro, Instituto Politécnico - Nova Friburgo, RJ, Brasil*
[2]*Instituto Politécnico, Universidade do Estado do Rio de Janeiro, Nova Friburgo, Brasil*
[†]**Autor correspondente:** camila.saporetti@iprj.uerj.br

Resumo

O petróleo e o gás são as principais fontes de energia primária no mundo. A partir desses recursos, obtêm-se derivados e petroquímicos que alimentam a produção de energia, serviços e diversos produtos. Entre as etapas cruciais da produção de petróleo estão a classificação dos reservatórios, a perfuração e a análise dos dados geológicos para determinar a viabilidade da extração. No entanto, esses processos costumam ser feitos manualmente por especialistas ou por métodos que são caros, imprecisos e demorados. Neste contexto, este trabalho tem o objetivo de classificar litologias e prever a taxa de carbono orgânico total por meio da aplicação de técnicas de aprendizado de máquina, empregando algoritmo genético com busca exaustiva para otimização dos métodos de regressão/classificação. A base de dados utilizada é referente a um poço do Campo Marlim, Bacia de Campos. Os resultados mostram que o Extreme Gradient Boosting (XGB) obteve bom desempenho nos experimentos realizados, com média de acurácia=0,941 e RMSE = 0,150 no conjunto de testes, sendo uma alternativa para auxiliar especialistas na tarefa de classificação de litologias e predição de taxa de carbono total.

Palavras-chave

Aprendizado de Máquina • Algoritmo Genético • Litologia • Carbono Orgânico Total

Abstract

Oil and gas are the main sources of primary energy in the world. From these resources, derivatives and petrochemicals are obtained that feed the production of energy, services and various products. Among the crucial stages of oil production are the classification of reservoirs, drilling and analysis of geological data to determine the feasibility of extraction. However, these processes are often done manually by experts or using methods that are expensive, inaccurate and time-consuming. In this context, this work aims to classify lithologies and predict the total organic carbon rate through the application of machine learning techniques, employing a genetic algorithm with exhaustive search to optimize regression/classification methods. The database used refers to a well in Campo Marlim, Campos Basin. The results show that Extreme Gradient Boosting (XGB) performed well in the experiments carried out, with average accuracy = 0.941 and RMSE = 0.150 in the test set, being an alternative to assist specialists in the task of lithology classification and rate prediction. total carbon.

☆Este artigo é uma versão estendida do trabalho apresentado no XXVII ENMC Encontro Nacional de Modelagem Computacional e XV ECTM Encontro de Ciência e Tecnologia de Materiais, ocorridos em Ilhéus – BA, de 1 a 4 de outubro de 2024.

Keywords

Machine Learning • Genetic Algorithm • Lithology • Total Organic Carbon

1 Introdução

O petróleo é uma mistura complexa de hidrocarbonetos, contendo também pequenas quantidades de outros compostos químicos [1]. Analisar esses componentes, entender suas interações e avaliar seu impacto na produção são etapas cruciais para determinar o potencial de produção de um poço [2]. Neste contexto, existe um grupo de áreas dentro da engenharia de petróleo dedicadas à utilização de ferramentas da matemática aplicada para que se obtenha um maior conhecimento das propriedades e das condições de escoamento nos reservatórios portadores de hidrocarbonetos [3]. Deve-se destacar, ainda, o papel fundamental das aplicações computacionais na construção de ferramentas úteis nos estudos de reservatórios [4].

Um aspecto crucial na caracterização de um reservatório é a litologia. A partir dos perfis petrofísicos coletados nos poços, é possível compreender o comportamento de um campo específico. A descrição das rochas, baseada em características como cor, composição mineralógica e tamanho dos grãos, é organizada em classes litológicas. Com esse entendimento, é possível avaliar o potencial e a heterogeneidade do reservatório [5].

A análise manual de litologias de perfil de poço é um procedimento trabalhoso que envolve um tempo gasto considerável por um especialista competente, mesmo quando auxiliado por métodos gráficos [6]. O problema torna-se particularmente desafiador à medida que aumenta o número de perfis de poços a serem avaliados. Portanto, seria útil automatizar o processo de caracterização de reservatórios.

Na avaliação de rochas geradoras de petróleo, o Carbono Orgânico Total (TOC) é um indicador-chave para a determinação do teor de hidrocarbonetos. A previsão precisa do TOC é essencial para a exploração e o desenvolvimento bem-sucedido dos recursos de petróleo e gás [7].

O TOC é uma maneira para qualificar a capacidade de geração da rocha geradora. A determinação precisa do TOC em amostras de solo e sedimentos é fundamental para a indústria de exploração de hidrocarbonetos, fornecendo informações essenciais sobre a presença e qualidade da matéria orgânica.

Os métodos mais utilizados para calcular o TOC são análises geoquímicas, realizadas em laboratório. Para isso, tornam-se necessários fragmentos de rocha ou mesmo testemunhos, aumentando os custos de exploração. Pesquisas que apresentam abordagens para estimar o TOC a partir de dados principais têm sido cada vez mais relatadas na literatura.

Diante do exposto, este trabalho busca aplicar métodos de aprendizado de máquina para classificar litologia e realizar a predição de TOC. Dessa forma, será possível auxiliar o processo de caracterização de reservatórios de petróleo diminuindo o tempo gasto nas análises.

A definição dos parâmetros ideais para maximizar o desempenho dos métodos de aprendizado de máquina é um problema comum. Para resolver essa questão, o Algoritmo Genético será empregado para otimizar os modelos, buscando encontrar os melhores parâmetros e melhorar a qualidade das estimativas.

O artigo está dividido da seguinte maneira: a Seção 2 trata dos trabalhos relacionados, a Seção 3 apresenta os dados utilizados e a metodologia empregada, a Seção 4 explora os resultados e a Seção 5 fornece a conclusão.

2 TRABALHOS RELACIONADOS

A previsão do Carbono Orgânico Total (TOC) e classificação de litologias são fundamentais para avaliar a capacidade de geração de hidrocarbonetos das rochas geradoras. Na literatura, diversas abordagens têm sido exploradas para tornais tais processos mais rápidos e com melhores desempenhos.

Yang et al. [8] aplicaram técnicas de transformada wavelet e agrupamento K-means modificado para classificar rochas metamórficas do Principal Furo Científico Continental Chinês (CCSD-MH). Os resultados mostraram maior precisão na identificação estratigráfica, destacando a eficácia dessa abordagem para melhorar a classificação de rochas metamórficas.

Elkatatny [9] propôs um método eficiente para estimar o teor de carbono orgânico total em reservatórios de folhelho utilizando registros petrofísicos. Com um modelo SaDE-ANN otimizado, alcançou alta precisão na predição do TOC usando dados como raios gama, tempo de compressão, resistividade e densidade bulk. A nova correlação empírica desenvolvida superou significativamente modelos anteriores, reduzindo os erros percentuais absolutos médios em até 67%.

Xie et al. [10] avaliaram cinco métodos de aprendizado de máquina (Naïve Bayes, Máquina de Vetores de Suporte, Rede Neural Artificial, Floresta Aleatória e Gradient Tree Boosting) usando dados dos campos de gás Daniudui e Hangjinqi. O estudo utilizou otimização de hiperparâmetros e validação cruzada para determinar o melhor modelo. Os resultados indicam que os métodos de ensemble apresentam menor erro de previsão e maior precisão na classificação da litologia, mesmo em classes de arenito.

Saporetti et al. [11] utilizaram seis métodos de aprendizado de máquina com técnicas de balanceamento de dados para classificar dados da Bacia de South Provence. Os resultados indicam que o balanceamento melhorou o desempenho dos classificadores e a seleção de modelos otimizou os parâmetros. A ferramenta computacional desenvolvida ajuda na identificação das heterogeneidades dos reservatórios.

O estudo de Asante-Okyere et al. [12] introduziu um modelo avançado de rede neural convolucional (CNN), o MWL-CNN, que integra dados de composição mineral do xisto e registros geofísicos de poços para previsão de TOC. Os resultados mostraram que a inclusão da composição mineral, especialmente componentes como feldspato e pirita, melhorou significativamente a precisão do modelo em comparação com abordagens baseadas apenas em registros de poços (WL-CNN).

Saporetti et al. [13] destacaram a importância da análise de TOC na exploração de intervalos geradores de hidrocarbonetos. Este estudo previu TOC utilizando uma abordagem híbrida, integrando modelos de aprendizado de máquina com o Algoritmo de Otimização Grey Wolf para ajustar parâmetros. A metodologia, avaliada com amostras do campo de gás de xisto YuDongNan, demonstrou que métodos de aprendizado de máquina assistidos por algoritmos evolutivos podem estimar TOC com precisão.

Silva et al. [14] propuseram uma abordagem para a previsão de TOC utilizando redes neurais convolucionais (CNNs) otimizadas por evolução diferencial. O estudo utilizou parâmetros selecionados por metaheurísticas e validação cruzada para melhorar a flexibilidade do modelo. A abordagem foi validada com amostras de várias bacias sedimentares, demonstrando o potencial das CNNs para prever concentrações de TOC de maneira eficiente e precisa.

Percebe-se que aplicar métodos de aprendizagem de máquina é algo promissor e empregar meta-heurística vai auxiliar a encontrar o melhor modelo, possibilitando realizar a previsão/classificação com um melhor desempenho. Então, objetiva-se avaliar o uso do Algoritmo Genético com busca exaustiva para encontrar os melhores métodos para classificar litologias e prever a taxa de Carbono Orgânico Total de um poço do Campo de Marlim.

3 MATERIAIS E MÉTODOS

3.1 Bases de Dados

O foco deste estudo é a área do Campo de Marlim, situado na região nordeste da Bacia de Campos, aproximadamente 110 km a leste do Cabo de São Tomé, na costa do Rio de Janeiro, com uma extensão total de $257,6km^2$.

Os dados petrofísicos de poços são disponibilizados pela Agência Nacional do Petróleo, Gás Natural e Biocombustíveis do Brasil (ANP), possuindo informações de 309 amostras e 12 características, que são Raios Gama (GR), Neutrônico (NPHI), Sônico (DT), Diâmetro de perfuração (CALI), Perfil de densidade (DRHO), Densidade (RHOB), Fator fotoelétrico (PEF), Caliper (CALI), Resistividade Profunda (ILD), Resistividade média (ILM), Resistividade microesférica (SFLA e SFLU) e Potencial Espontâneo (SP) além dos valores de TOC e as classes litológicas que são divididas em Arenito, Marga e Argilito.

3.2 Validação Cruzada

A validação cruzada k-fold (k-fold cross-validation) é um procedimento de divisão dos objetos nos conjuntos de treinamento e teste, em que cada objeto é utilizado uma única vez em um dos k conjuntos de teste e (k-1) vezes em um dos k conjuntos de treinamento [15]. Esse processo é repetido k vezes, utilizando em cada ciclo uma partição diferente para o teste, sendo o desempenho final dado pela média dos desempenhos observados sobre cada subconjunto de teste [16].

3.3 Métodos

Para o processo de modelagem computacional dos dados foram aplicados os seguintes algoritmos supervisionados de Aprendizado de Máquina que podem ser utilizados tanto em regressão como em classificação: K-Nearest Neighbors (KNN), Extreme Learning Machine (ELM), Support Vector Machines (SVM) e Extreme Gradient Boosting (XGB).

O ELM é uma rede neural artificial feedforward com apenas uma camada oculta, com pesos de conexão de entrada escolhidos aleatoriamente [17, 18]. A saída do ELM é descrita como

$$\hat{y} = \sum_{i=1}^{L} \beta_i G\left(\alpha_i \mathbf{x} + b_i\right) \tag{1}$$

onde $\{(x_i, y_i), x_i \in \mathfrak{R}^n, y_i \in \mathfrak{R}^1, i = 1, 2, \cdots, N\}$ são as amostras de treinamento, L o número de neurônios ocultos, $\{\beta_i, i = 1, 2, \cdots, N\}$ os pesos de saída, G a função de ativação, $\{\alpha_i, i = 1, 2, \cdots, N\}$ é o vetor de pesos, b_i é o bias para o nó oculto i, e $\hat{y}$ é a saída predita.

A Eq. (1) pode ser estruturada como $\mathbf{H}\beta = \mathbf{T}$, onde $H_{ij} = G(\alpha_j, b_j, \mathbf{x}_i)$ e $T_i = y_i$. A função objetivo quadrática $\sum_{i=1}^{N} \parallel \hat{y}_i - y_i \parallel = 0$ é minimizada usando mínimos quadrados, e o vetor de peso de saída é dado por $\beta = \left(\mathbf{H}^T\mathbf{H}\right)^{-1}\mathbf{H}^T\mathbf{T}$.

Máquina de Vetores de Suporte (SVM) é um modelo de aprendizado de máquina capaz de fazer classificações lineares e não lineares [19, 20] e regressão [21]. O modelo linear do classificador SVM prediz a classe de uma instância nova $\mathbf{x}$ calculando a função de decisão $\mathbf{w}^T\mathbf{x} + b$, onde b é o bias e $\mathbf{w}$ é o vetor de pesos das características. Assim, a saída é classificada de acordo com a Eq. (2)

$$\hat{y} = \begin{cases} 0 & , se\ \mathbf{w}^T\mathbf{x} < 0 \\ 1 & , se\ \mathbf{w}^T\mathbf{x} \geq 0 \end{cases} . \tag{2}$$

Portanto, treinar um classificador SVM linear significa encontrar os valores de $\mathbf{w}$ e b que fazem com que a margem ao redor da fronteira de decisão seja a mais ampla possível, ao passo que a margem seja rígida (evita as violações de margem) ou seja suave (restringindo as violações de margem), sendo controlado pelo parâmetro C na Eq. (3). Para classificação SVM não linear é empregado o truque do kernel. Já para regressão SVM, em vez de tentar ajustar a maior largura de margem possível entre as duas classes enquanto se restringe as violações de margem, tenta ajustar o maior número possível de instâncias entre as margens enquanto restringe a margem de violações (ou seja, das instâncias fora da "rua" entre as margens).

$$\begin{array}{ll} minimiza & \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{m}\zeta^i \\ sujeita\ a & t^{(i)}(\mathbf{w}^T\mathbf{x}^{(i)} + b) \geq 1 - \zeta^i\ em\ \zeta^i \geq 0,\ i = 1, \cdots, m \end{array} \tag{3}$$

onde $\zeta^{(i)} \geq 0$ é uma variável que calcula o quanto a instância i pode violar a margem, e o hiperparâmetro C permite definir a troca entre um classificador linear com margem rígida ($C = 0$) e um classificador linear de margem suave ($C = 1$). Importante ressaltar que, como o algoritmo SVM usa internamente o cálculo de distância, deve ser feito o escalonamento dos atributos preditivos no pré-processamento.

KNN foi desenvolvido pela primeira vez por Fix e Hodges [22], e posteriormente expandido por Cover e Hart [23] é um algoritmo baseado em proximidade que usa distância euclidiana para avaliar a proximidade entre cada par de objetos, assumindo que quanto menor for a distância entre dois objetos mais semelhantes eles são [15]. Escolhe-se um objeto aleatoriamente, a partir disso analisa a classe dos K vizinhos mais próximos, a classe que aparece na maioria dos K vizinhos é atribuída ao objeto. O KNN apesar de ser um método simples é propenso ao overfitting pelos seguintes fatores: é sensível ao ruído nos dados de treinamento, podendo afetar a predição em novas amostras, em espaços de alta dimensionalidade, a distância entre os pontos se torna menos significativa podendo levar a uma percepção errada de similaridade entre pontos e se os dados de treinamento são muito desbalanceados ou possuem uma distribuição irregular, o KNN pode se ajustar excessivamente a essas características específicas dos dados de treinamento, prejudicando sua capacidade de generalização [24].

O algoritmo XGB desenvolvido por Chen e Guestrin [25] é do tipo ensemble, combinando modelos sequencialmente, de modo que é atribuída uma maior importância ao aprendizado de objetos que o modelo anterior não conseguiu apresentar uma boa predição [26]. É baseado no princípio do gradiente descendente, um algoritmo de otimização onde novas árvores são geradas com base nas anteriores, visando reduzir a função objetivo dada pela Eq. (4) a um menor valor possível [27].

$$Obj = \sum_{j=1}^{T}\left[G_j\omega_j + \frac{1}{2}\left(H_j + \lambda\right)\omega_j^2\right] + \gamma T \tag{4}$$

onde T é o número de folhas, G é a soma do gradiente da função de perda, H é a soma do Hessiano da função de perda, ω é o vetor das pontuações (scores) nas folhas, λ e γ representam os coeficientes de penalidade.

Na abordagem ensemble do tipo boosting, os modelos são induzidos sequencialmente (a saída gerada por um modelo é recebida como entrada por outro modelo). Para a indução de cada novo modelo, é atribuída uma maior importância ao aprendizado de objetos que o modelo anterior não conseguiu apresentar uma boa predição [26]. Assim, no modelo XGB as árvores de decisão são adicionadas sequencialmente como aprendizes fracos (weak learners) produzindo um aprendiz forte (strong learner), sendo cada nova árvore treinada para corrigir os erros cometidos usando a função objetivo dada pela Eq. (4).

3.4 Otimização evolutiva sobre hiperparâmetros

O GASearchCV é uma otimização evolutiva sobre hiperparâmetros, tendo em vista os scores de validação cruzada, que melhora a precisão da previsão do algoritmo genético usando testes aleatórios de seu hiperparâmetro inicial [28]. Pode ser usado tanto para problemas de regressão quanto de classificação. A vantagem da implementação

usando busca em grade dos hiperparâmetros é que consegue-se definir quais hiperparâmetros são inteiros, quais são categóricos e quais são reais de forma que a variação é discreta ou contínua de acordo com o tipo. Dessa, forma não precisa fazer a busca contínua em todos os parâmetros e realizar uma transformação para inteiro e posterior atribuição a dados categóricos.

O método de validação cruzada utilizado foi o K-Fold com k=5, tamanho da população = 20, nº de gerações = 30, probabilidade de crossover = 0,9, probabilidade de mutação = 0,08, scoring = acurácia para classificação de litologia e scoring = raiz do erro quadrático médio para predição de TOC. A descrição dos modelos com os hiperparâmetros que foram otimizados e as respectivas variações encontra-se na Tabela 1.

Tabela 1: Descrição dos modelos

Método	Parâmetros	Descrição	Configuração
ELM	*activation_func*	Função de ativação.	[identity, logistic, tanh, reLu]
	n_hidden	Número de neurônios.	[20, 150]
	alpha	Força de regularização.	[0.001, 10]
KNN	*n_neighbors*	Número de vizinhos mais próximos.	[2, 30]
	weights	Função de peso usada na previsão.	[uniform, distance]
SVM	*C*	Ajusta a penalidade dos erros na regressão/classificação.	[20, 200]
	gamma	Define até onde chega a influência de um único exemplo de treinamento.	[0.001, 0.1]
XGB	*n_estimators*	Número de árvores na floresta.	[5, 300]
	max_depth	Profundidade máxima.	[2, 10]
	learning_rate	Taxa de aprendizagem.	[0.001, 0.1]

3.5 Métricas

Para a avaliação do desempenho dos métodos utilizados para a predição de TOC, foram utilizadas as seguintes métricas: o coeficiente de determinação (R^2), o erro médio absoluto(MAE), o erro quadrático médio (MSE), raiz do erro quadrático médio (RMSE) e o erro relativo absoluto médio (MARE) . Onde y_i representa os valores medidos da variável dependente (TOC), $\hat{y}_i$ representa os valores preditos de TOC, $\bar{y}_i$ o valor médio de TOC e n o tamanho da amostra.

- $R^2 = \sum_{i=1}^{N}(y_i - \hat{y}_i)^2 / \sum_{i=1}^{N}(y_i - \bar{y})$
- $\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}|$
- $\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y})^2$
- $\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y})^2}$
- $\text{MARE} = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{y_i - \hat{y}}{y_i}\right|$

Para a classificação de litologia foram empregadas as seguintes métricas Acurácia (AC), F1, Recall (Re), Kappa e Precisão (Pr).

- $AC = \frac{1}{N}\sum_{i=1}^{N} I(f(x_i) = y_i)$
  onde, $f(x_i)$ é a classe predita as amostras de teste e y_i é a classe verdadeira dessas amostras. Considerando que $I(true) = 1$ e $I(false) = 0$.
- $F1 = \frac{2TP}{2TP + FP + FN}$
- $Re = \frac{TP}{TP + FN}$

Tabela 2: Média e Desvio Padrão das Métricas na Predição de TOC - Conjunto de Treinamento

Método	R^2	RMSE	MAE	MARE	MSE
ELM	0.238 (0.042)	0.175 (0.009)	0.136 (0.007)	0.477 (0.024)	0.031 (0.003)
KNN	**0.950** (0.151)	**0.014** (0.043)	**0.010** (0.031)	**0.035** (0.106)	**0.002** (0.006)
SVR	0.673 (0.081)	0.114 (0.017)	0.067 (0.016)	0.242 (0.055)	0.013 (0.004)
XGB	0.754 (0.107)	0.096 (0.028)	0.065 (0.018)	0.230 (0.065)	0.010 (0.004)

Tabela 3: Média e Desvio Padrão das Métricas na Predição de TOC - Conjunto de Teste

Método	R^2	RMSE	MAE	MARE	MSE
ELM	0.139 (0.121)	0.186 (0.029)	0.145 (0.013)	0.510 (0.054)	0.035 (0.012)
KNN	0.294 (0.138)	0.167 (0.026)	0.123 (0.011)	0.414 (0.039)	0.029 (0.009)
SVR	0.329 (0.110)	0.164 (0.028)	0.117 (0.012)	0.421 (0.040)	0.028 (0.010)
XGB	**0.422** (0.108)	**0.150** (0.025)	**0.103** (0.012)	**0.353** (0.038)	**0.023** (0.008)

- $Pr = \dfrac{TP}{TP + FP}$
  TP, FN e FP são o número de verdadeiros positivos negativos e falsos positivos.
- $KAPPA = \dfrac{P_o - P_E}{1 - P_E}$
  - $P_o = \dfrac{\text{nº de concordâncias}}{\text{nº de concordâncias} + \text{nº discordâncias}}$
  - $P_E = \sum_{i=1}^{N}(p_{i1} \times p_{i2})$
    onde N é o número de categorias, i é o índice de categorias, p_{i1} é a ocorrência da categoria de proporção i para o avaliador 1, p_{i2} é a ocorrência da categoria de proporção i para o avaliador 2.

4 RESULTADOS E DISCUSSÃO

Para a avaliação do desempenho dos métodos utilizados para a predição de TOC, foram utilizadas as seguintes métricas: o coeficiente de determinação (R^2), o erro médio absoluto (MAE), o erro quadrático médio (MSE), raiz do erro quadrático médio (RMSE) e o erro relativo absoluto médio (MARE). As Tabelas 2 e 3 apresentam os desempenhos do métodos utilizados na predição de TOC para o conjunto de treinamento e teste, respectivamente. Pode-se observar que o KNN obteve o melhor resultado no conjunto de treinamento ($R^2 = 0.950$, RMSE = 0.014), seguido do XGB ($R^2 = 0.754$, RMSE = 0.096). Porém no conjunto de teste o XGB ($R^2 = 0.422$, RMSE = 0.150) foi o método com melhor resultado e o KNN ($R^2 = 0.294$, RMSE = 0.167) o terceiro com melhor resultado. Este resultado sugere que houve um overfitting com o KNN, pois o desempenho dele no conjunto de treinamento foi muito superior ao de teste e não se destacou comparando aos outros métodos. Isso indica que o aprendizado do KNN ficou limitado ao conjunto de treinamento, não tendo a capacidade de aplicar o aprendizado em amostras desconhecidas.

A Figura 1 exibe os boxplots das métricas no conjunto de teste na predição de TOC nas 30 execuções realizadas. Nota-se que o XGB apresentou melhor desempenho se comparado aos outros métodos além de ter uma menor variação nos valores das métricas.

Para a classificação de litologia foram empregadas as seguintes métricas Acurácia (AC), F1, Recall (Re), Kappa e Precisão (Pr). As Tabelas 4 e 5 mostram os desempenhos do métodos utilizados na classificação de litologia para o conjunto de treinamento e teste, respectivamente. Nota-se que um comportamento similar a predição de TOC ocorre, pelo fato de se tratar dos mesmo dados de entrada porém com variável de saída diferente. O KNN obteve melhor resultado no conjunto de treino e o XGB no de teste e nesse caso também o indicativo de overfitting.

A Figura 2 exibe os boxplots das métricas no conjunto de teste na classificação de litologia nas 30 execuções realizadas. Nota-se que o XGB apresentou melhor desempenho se comparado aos outros métodos além de ter uma menor variação nos valores das métricas.

A Figura 3 mostra para cada modelo de aprendizagem de máquina, a distribuição dos parâmetros internos para predição de TOC. Pode-se observar que para o ELM o α varia entre aproximadamente 0,5 e 3,0, nƐ de neurônios em torno de 80 e 140 e a função de ativação que foi escolhida na maioria das iterações foi tangente hiperbólica. Para o KNN, o número de vizinhos variou entre 7 e 11, o weight foi distance em 29 iterações. No caso do SVR, o C variou

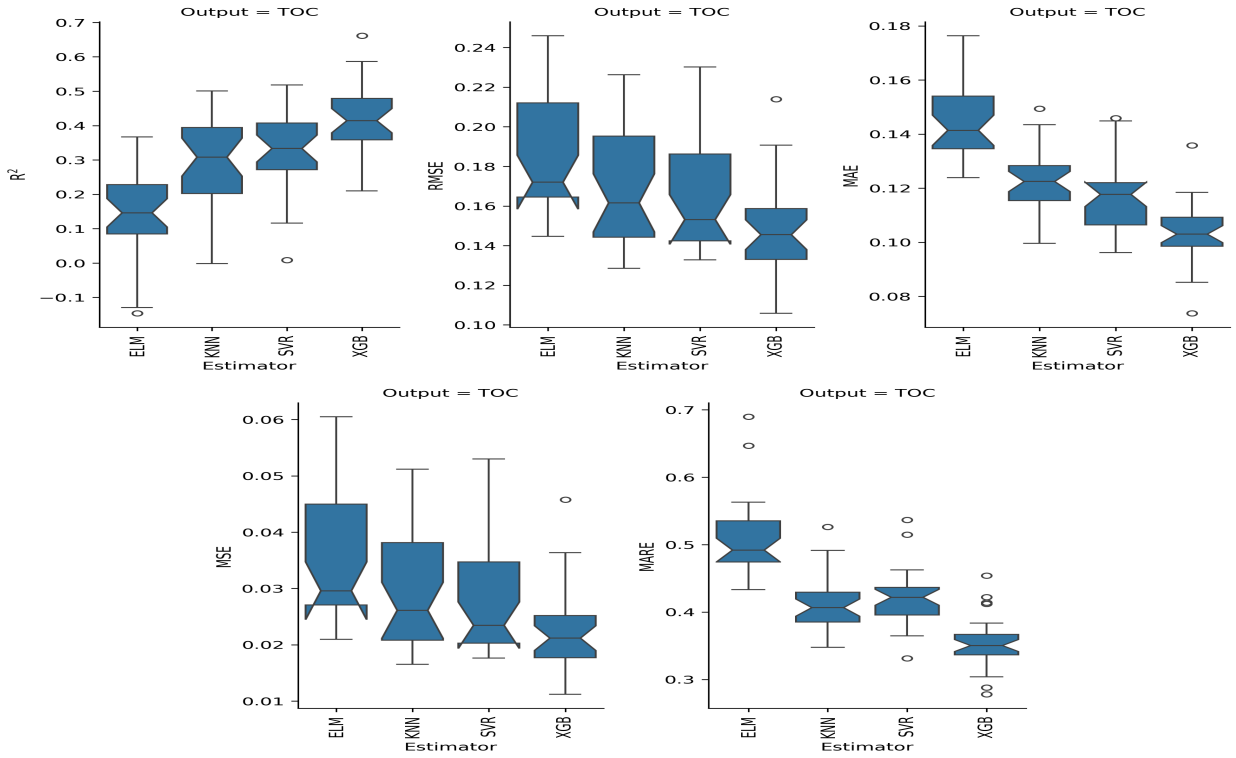

Figura 1: *Boxplot* para as métricas R^2, RMSE, MAE, MSE e MARE.

Tabela 4: Média e Desvio Padrão das Métricas na Classificação de Litologia - Conjunto de Treinamento

MÉTODO	ACURÁCIA	F1	RECALL	KAPPA	PRECISÃO
ELM	0.951 (0.034)	0.951 (0.035)	0.951 (0.034)	0.916 (0.060)	0.953 (0.032)
KNN	**1.000** (–)	**1.000** (–)	**1.000** (–)	**1.000** (–)	**1.000** (–)
SVM	0.940 (0.036)	0.939 (0.037)	0.940 (0.036)	0.897 (0.063)	0.943 (0.034)
XGB	0.996 (0.007)	0.996 (0.007)	0.996 (0.007)	0.994 (0.013)	0.996 (0.007)

Tabela 5: Média e Desvio Padrão das Métricas na Classificação de Litologia - Conjunto de Teste

MÉTODO	ACURÁCIA	F1	RECALL	KAPPA	PRECISÃO
ELM	0.827 (0.055)	0.802 (0.064)	0.827 (0.055)	0.683 (0.095)	0.835 (0.065)
KNN	0.820 (0.042)	0.820 (0.043)	0.820 (0.042)	0.683 (0.073)	0.830 (0.039)
SVM	0.858 (0.048)	0.856 (0.050)	0.858 (0.048)	0.749 (0.083)	0.866 (0.046)
XGB	**0.941** (0.029)	**0.941** (0.029)	**0.941** (0.029)	**0.896** (0.053)	**0.944** (0.028)

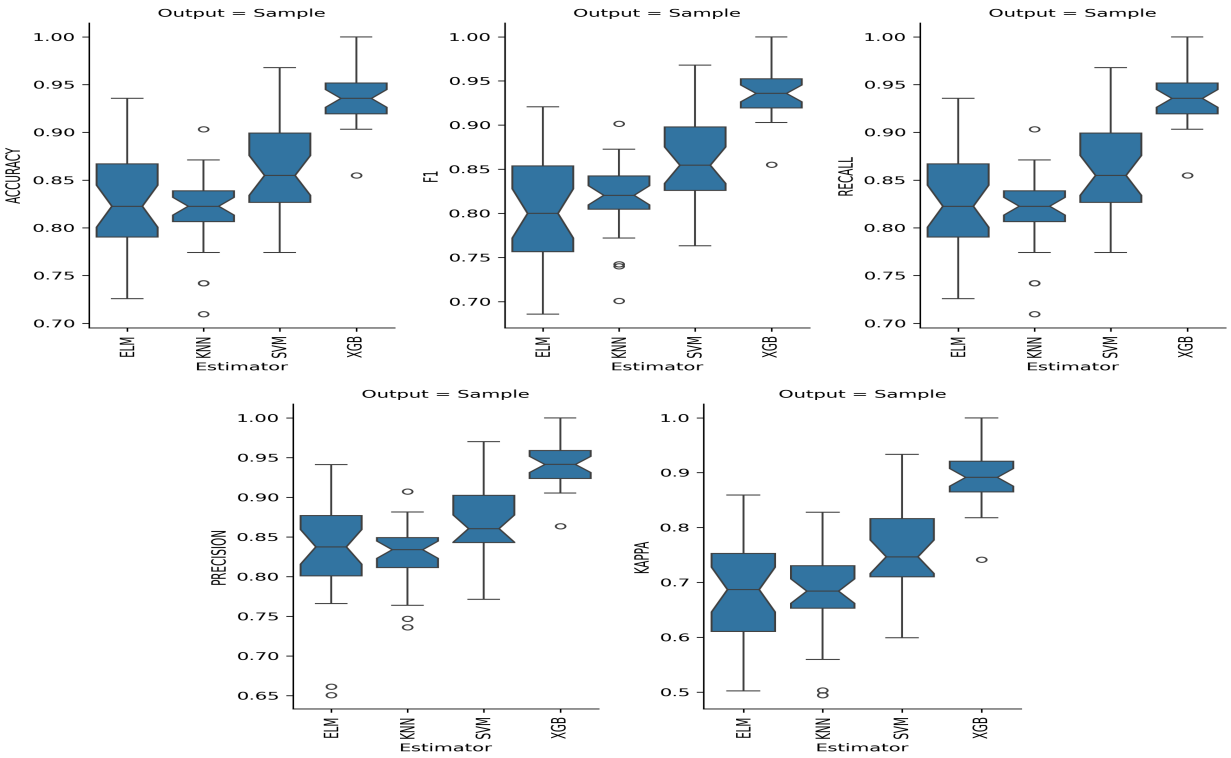

Figura 2: *Boxplot* para as métricas Acurácia, F1, Recall, Precision e Kappa .

de aproximadamente 0,5 a 2, o γ entre 0,05 e 0,1. Para o XGB, o parâmetro learning rate variou de 0,02 e 0,055, a profundidade máxima na maioria da iterações foi 2 e o número de estimadores de 90 a 250.

A Figura 4 apresenta para cada método, a distribuição dos parâmetros internos para classificação de litologia. Nota-se que para o ELM o α varia entre aproximadamente 1,9 e 5,8, nƐ de neurônios em torno de 100 e 130 e a função de ativação que foi escolhida na maioria das iterações foi ReLU. Para o KNN, o número de vizinhos 2 e 4 foram escolhidos em 10 iterações cada, o weight foi distance em 29 iterações. No caso do SVM, o C variou de aproximadamente 10 a 300, o γ entre 0 e 0,03. Para o XGB, o parâmetro learning rate variou de 0,03 e 0,07, a profundidade máxima foi 3 em 9 iterações e o número de estimadores de 150 a 250. Pode-se observar que para a classificação houve mais esforço tendo em vista os valores dos parâmetros mais altos em grande parte dos casos.

A Tabela 6 mostra os melhores modelos encontrados após 30 execuções independentes na predição de TOC e seus resultados no conjunto de teste. Observa-se que o melhor modelo foi referente ao XGB com $R^2 = 0.661$ e RMSE $= 0.106$.

Tabela 6: Melhores Modelos (TOC) - Conjunto de Teste

Modelo	Melhores Parâmetros	MSE	RMSE	MAE	MARE	R^2
ELM	no. neurons = 122, activation function = tanh alpha=2.67	0.021	0.145	0.124	0.467	0.367
KNN	no. neighbors = 6, weight=distance	0.017	0.129	0.100	0.365	0.500
SVR	C = 0.245, gamma = 0.077	0.018	0.133	0.101	0.413	0.467
XGB	no. estimators=299, max. depth = 3 learning rate = 0.020	0.011	0.106	0.074	0.278	0.661

A Tabela 7 mostra os melhores modelos encontrados após 30 execuções independentes na na classificação de litologia e seus resultados no conjunto de teste. Observa-se que o melhor modelo foi referente ao XGB com acurácia $= 1.000$ e Kappa $= 1.000$, mostrando que o nível de concordância é perfeito.

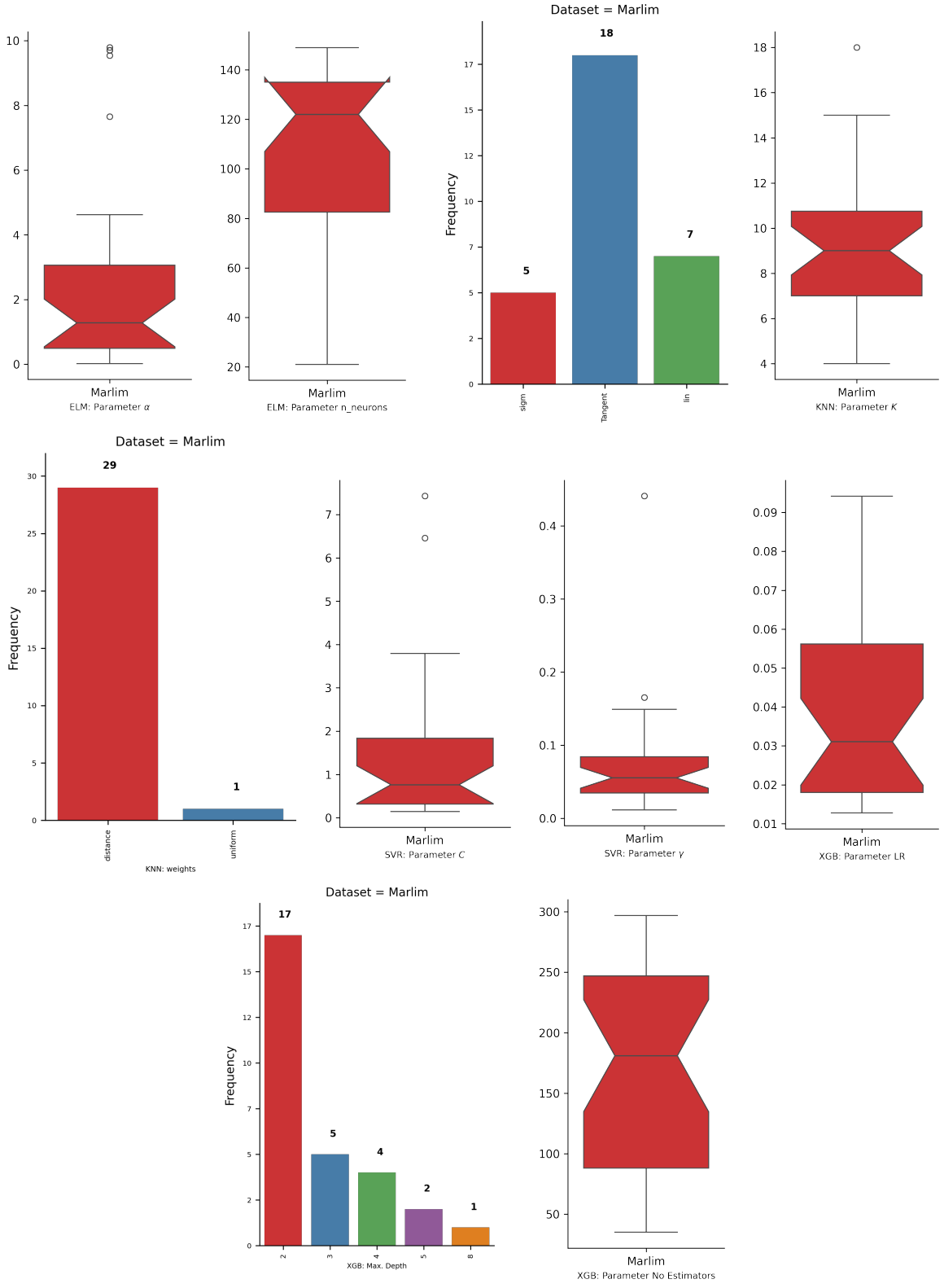

Figura 3: Distribuição dos parâmetros para ELM, KNN, SVR e XGB em 30 execuções independentes - Predição TOC.

5 Conclusões

Este estudo avaliou o uso de algoritmo genético combinado com quatro modelos de aprendizado de máquinas para um problema de previsão de TOC e classificação de litologia com dados coletados no campo de Marlim, bacia de

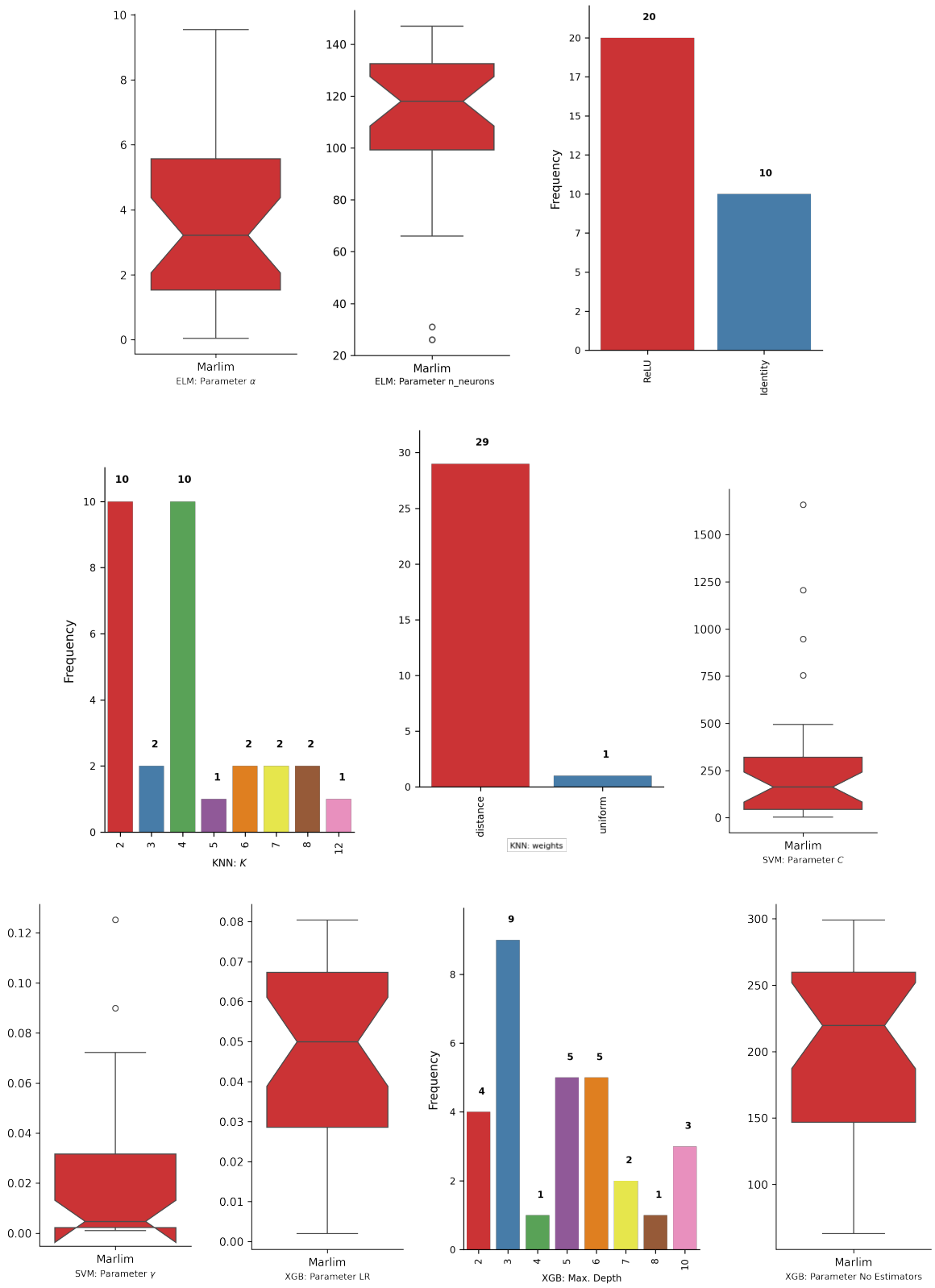

Figura 4: Distribuição dos parâmetros para ELM, KNN, SVR e XGB em 30 execuções independentes - Classificação Litologia.

Campos.

A abordagem consistiu em formular o problema de seleção de modelos como um problema de otimização bus-

Tabela 7: Melhores Modelos (Litologia) - Conjunto de Teste

Modelo	Melhores Parâmetros	Acurácia	F1	Precisão	Recall	Kappa
ELM	no. neurons = 144, activation function = relu, alpha=7.68	0.935	0.921	0.941	0.935	0.859
KNN	no. neighbors = 4, weight=distance	0.903	0.901	0.907	0.903	0.828
SVC	C = 176.198, gamma = 0.968	0.968	0.133	0.970	0.968	0.968
XGB	no. estimators=198, max. depth = 2 learning rate = 0.034	1.000	1.000	1.000	1.000	1.000

cando regiões de mínimos/máximos locais e um hiperespaço de parâmetros considerando o erro quadrático médio/acurácia como função objetivo. Ao final do processo de otimização, foram calculadas as médias de desempenho para comparar a eficácia dos algoritmos populacionais no ajuste dos parâmetros do modelo de aprendizado de máquina.

O método que obteve melhor resultado foi o XGB tanto para predição de TOC quanto para classificação de litologia, mostrando ser uma possibilidade para automatizar tais processos e auxiliar na tomada de decisão de especialistas, reduzindo tempo e custo.

Agradecimentos

O presente trabalho foi realizado com o apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001. Os autores agradecem a Fundação Carlos Chagas Filho de Amparo à Pesquisa do Estado do Rio de Janeiro (FAPERJ) pelo suporte financeiro através 10.432/2024-APQ1.

Referências

[1] N. Ezekwe, *Petroleum Reservoir Engineering Practice*. Westford, USA: Prentice Hall, 2010.

[2] N. Al-Mohannadi, "Simulation of horizontal well tests," Tese de doutorado, Colorado School of Mines, Golden, USA, 2004.

[3] A. J. Rosa, R. S. Carvalho, e J. A. D. Xavier, *Engenharia de Reservatórios de Petróleo*. Interciência, 2006.

[4] M. Mohammadpour, H. Roshan, M. Arashpour, e H. Masoumi, "Machine learning assisted Kriging to capture spatial variability in petrophysical property modelling," *Marine and Petroleum Geology*, vol. 167, p. 106967, 2024. Disponível em: https://doi.org/10.1016/j.marpetgeo.2024.106967

[5] C. M. Saporetti, L. Goliatt, e E. Pereira, "Neural network boosted with differential evolution for lithology identification based on well logs information," *Earth Science Informatics*, vol. 14, no. 1, pp. 133–140, 2021. Disponível em: https://doi.org/10.1007/s12145-020-00533-x

[6] E. M. Vasini, A. Battistelli, P. Berry, S. Bonduà, V. Bortolotti, C. Cormio, e L. Pan, "Interpretation of production tests in geothermal wells with T2Well-EWASG," *Geothermics*, vol. 73, pp. 158–167, 2018. Disponível em: https://doi.org/10.1016/j.geothermics.2017.06.005

[7] G. van Graas, T. Viets, J. De Leeuw, e P. Schenck, "A study of the soluble and insoluble organic matter from the Livello Bonarelli, a cretaceous black shale deposit in the Central Apennines, Italy," *Geochimica et Cosmochimica Acta*, vol. 47, no. 6, pp. 1051–1059, 1983. Disponível em: https://doi.org/10.1016/0016-7037(83)90235-1

[8] H. Yang, H. Pan, H. Ma, A. A. Konaté, J. Yao, e B. Guo, "Performance of the synergetic wavelet transform and modified K-means clustering in lithology classification using nuclear log," *Journal of Petroleum Science and Engineering*, vol. 144, pp. 1–9, 2016. Disponível em: https://doi.org/10.1016/j.petrol.2016.02.031

[9] S. Elkatatny, "A self-adaptive artificial neural network technique to predict total organic carbon (TOC) based on well logs," *Arabian Journal of Science and Engineering*, vol. 44, pp. 6127–6137, 2018. Disponível em: https://doi.org/10.1007/s13369-018-3672-6

[10] Y. Xie, C. Zhu, W. Zhou, Z. Li, X. Liu, e M. Tu, "Evaluation of machine learning methods for formation lithology identification: A comparison of tuning processes and model performances," *Journal of Petroleum Science and Engineering*, vol. 160, pp. 182–193, 2018. Disponível em: https://doi.org/10.1016/j.petrol.2017.10.028

[11] C. M. Saporetti, L. G. da Fonseca, E. Pereira, e L. C. de Oliveira, "Machine learning approaches for petrographic classification of carbonate-siliciclastic rocks using well logs and textural information," *Journal of Applied Geophysics*, vol. 155, pp. 217–225, 2018. Disponível em: https://doi.org/10.1016/j.jappgeo.2018.06.012

[12] S. Asante-Okyere, Y. Y. Ziggah, e S. A. Marfo, "Improved total organic carbon convolutional neural network model based on mineralogy and geophysical well log data," *Unconventional Resources*, vol. 1, pp. 1–8, 2021. Disponível em: https://doi.org/10.1016/j.uncres.2021.04.001

[13] C. Saporetti, D. Fonseca, L. Oliveira, E. Pereira, e L. Goliatt, "Hybrid machine learning models for estimating total organic carbon from mineral constituents in core samples of shale gas fields," *Marine and Petroleum Geology*, vol. 143, p. 105783, 2022. Disponível em: https://doi.org/10.1016/j.marpetgeo.2022.105783

[14] R. O. Silva, C. M. Saporetti, Z. M. Yaseen, E. Pereira, e L. Goliatt, "An approach for total organic carbon prediction using convolutional neural networks optimized by differential evolution," *Neural Computing and Applications*, vol. 35, pp. 20 803–20 817, 2023. Disponível em: https://doi.org/10.1007/s00521-023-08865-7

[15] A. de Carvalho, A. Menezes, e R. Bonidia, *Ciência de Dados - Fundamentos e Aplicações*. LTC, 2024. Disponível em: https://books.google.com.br/books?id=HNSn0AEACAAJ

[16] K. Faceli, A. C. Lorena, J. Gama, T. A. d. Almeida, e A. C. P. L. F. Carvalho, *Inteligência artificial: uma abordagem de aprendizado de máquina*. LTC, 2021.

[17] G.-B. Huang, Q.-Y. Zhu, e C.-K. Siew, "Extreme learning machine: theory and applications," *Neurocomputing*, vol. 70, no. 1-3, pp. 489–501, 2006. Disponível em: https://doi.org/10.1016/j.neucom.2005.12.126

[18] ——, "Extreme learning machine: a new learning scheme of feedforward neural networks," em *2004 IEEE International Joint Conference on Neural Networks (IEEE Cat. No. 04CH37541)*, vol. 2. IEEE, 2004, pp. 985–990. Disponível em: https://doi.org/10.1109/IJCNN.2004.1380068

[19] C. Cortes e V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, pp. 273–297, 1995. Disponível em: https://doi.org/10.1007/BF00994018

[20] V. N. Vapnik, *The Nature of Statistical Learning Theory*. Springer New York, 2000. Disponível em: https://doi.org/10.1007/978-1-4757-3264-1

[21] H. Drucker, C. J. Burges, L. Kaufman, A. Smola, e V. Vapnik, "Support vector regression machines," *Advances in Neural Information Processing Systems*, vol. 9, 1996.

[22] E. Fix e J. L. Hodges, "Discriminatory analysis, nonparametric discrimination," 1951.

[23] T. Cover e P. Hart, "Nearest neighbor pattern classification," *IEEE Transactions on Information Theory*, vol. 13, no. 1, pp. 21–27, 1967. Disponível em: https://doi.org/10.1109/TIT.1967.1053964

[24] R. Souza, R. Lotufo, e L. Rittner, "A comparison between optimum-path forest and k-nearest neighbors classifiers," em *2012 25th SIBGRAPI Conference on Graphics, Patterns and Images*. IEEE, 2012, pp. 260–267. Disponível em: https://doi.org/10.1109/SIBGRAPI.2012.43

[25] T. Chen e C. Guestrin, "Xgboost: A scalable tree boosting system," em *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 785–794. Disponível em: https://doi.org/10.1145/2939672.2939785

[26] R. E. Schapire, "The strength of weak learnability," *Machine Learning*, vol. 5, pp. 197–227, 1990. Disponível em: https://doi.org/10.1007/BF00116037

[27] F. R. Bione, I. M. Venancio, T. P. Santos, A. L. Belem, B. R. Rangel, I. V. Souza, A. L. Spigolon, e A. L. S. Albuquerque, "Estimating total organic carbon of potential source rocks in the Espírito Santo basin, SE Brazil, using XGBoost," *Marine and Petroleum Geology*, vol. 162, p. 106765, 2024. Disponível em: https://doi.org/10.2139/ssrn.4631704

[28] R. A. Gómez, *GASearchCV*, 2021. Disponível em: https://sklearn-genetic-opt.readthedocs.io/en/stable/