APLICAÇÃO DE TÉCNICAS DE APRENDIZADO DE MÁQUINA NA CLASSIFICAÇÃO LITOLÓGICA DE POÇOS DE PETRÓLEO

Juliana da Costa Cabral[1] - juliana.costa@iprj.uerj.br
Clovis Antonio da Silva[1] - casilva@iprj.uerj.br
Camila Martins Saporetti[2] - camila.saporetti@iprj.uerj.br
[1]Pós-Graduação em Modelagem Computacional, Universidade do Estado do Rio de Janeiro, Instituto Politécnico - Nova Friburgo, RJ, Brasil
[2]Departamento de Modelagem Computacional, Universidade do Estado do Rio de Janeiro, Instituto Politécnico - Nova Friburgo, RJ, Brasil

Resumo. *A classificação litológica é uma etapa fundamental na caracterização de reservatórios de petróleo, pois permite identificar os tipos de rochas e suas propriedades físicas e minerais, o que contribui para a análise da capacidade de armazenamento e do escoamento de hidrocarbonetos. Métodos tradicionais, como a análise manual de perfis geofísicos, ainda são utilizados, mas enfrentam limitações por dependerem da interpretação humana e não serem eficientes para grandes volumes de dados. Nesse cenário, o uso de técnicas de aprendizado de máquina tem se destacado por permitir a detecção de padrões complexos de forma automatizada e precisa. Este trabalho propõe uma abordagem computacional para a classificação litológica por meio de algoritmos de aprendizado de máquina supervisionados combinados com a técnica de extração de características K-Means Featurizer. Empregou-se dados de três poços do Campo de Marlim, localizado na Bacia de Campos. Os modelos foram otimizados com Grid Search e validados usando validação cruzada K-fold. A avaliação foi feita com base nas métricas de acurácia, precisão, recall e F1-score. O XGB obteve o melhor desempenho, alcançando acurácia de 0,910. Os resultados evidenciam o potencial da combinação entre métodos de agrupamento e modelos supervisionados na melhoria da classificação litológica em contextos sedimentares de difícil interpretação.*

Palavras-chave: *Classificação Litológica, Aprendizado de Máquina, K-Means Featurizer, Reservatórios de Petróleo*

1. INTRODUÇÃO

A classificação litológica consiste em organizar as rochas em diferentes categorias ou classes, com base nas suas propriedades físicas e mineralógicas (Saporetti et al., 2021). Por meio dessa categorização, é possível interpretar as variações do reservatório e prever seu desempenho, uma vez que diferentes litologias apresentam comportamentos distintos frente à presença e ao fluxo de hidrocarbonetos (Hu et al., 2013; Chen et al., 2010).

Tradicionalmente, a litologia é identificada por métodos diretos, como a observação macroscópica de testemunhos de rochas ou a análise microscópica de lâminas finas, que focam em aspectos como cor, estrutura mineral e textura das rochas (Hu et al., 2013; Chen et al., 2010). Além disso, métodos experimentais baseados em propriedades geoquímicas e geofísicas, como densidade, magnetismo, condutividade elétrica e composição mineralógica, também são amplamente empregados para caracterizar as rochas (Fu et al., 2017; Xia et al., 2010). Entretanto, esses métodos são caros, exigem alta especialização técnica, equipamentos avançados e são suscetíveis a vieses subjetivos devido à dependência da interpretação humana.

Nesse contexto, os métodos indiretos surgem como uma alternativa eficaz para superar algumas das limitações associadas aos procedimentos tradicionais. Eles se baseiam na aplicação de técnicas de perfilagem (*well logging*), que permitem registrar as propriedades físicas e geológicas das formações atravessadas por um poço de petróleo, ao longo de suas diferentes camadas. Essas técnicas produzem dados essenciais para a caracterização das formações, sem a necessidade de extração física de amostras (Saporetti et al., 2021; Abbas and Rasool, 2024).

A perfilagem gera um volume expressivo de dados, com milhares de amostras coletadas ao longo do poço e múltiplos atributos petrofísicos por ponto. Para lidar com essa quantidade de informações, utilizam-se métodos automatizados, como técnicas de aprendizado de máquina, que permitem identificar padrões e correlações não evidentes aos métodos tradicionais, aumentando a precisão e reduzindo o tempo de análise (Ambagtsheer et al. 2020; Ibrahim et al. 2020).

Na literatura, diversas estratégias têm sido desenvolvidas para acelerar esses processos e aprimorar seu desempenho, contribuindo para decisões mais precisas na indústria do petróleo. Hou et al. (2023) aplicaram modelos de aprendizado supervisionado, incluindo Perceptron de Múltiplas Camadas, Máquina de Vetores de Suporte, Extreme Gradient Boosting e Floresta Aleatória, para classificar litofácies no folhelho da formação Gulong, utilizando dados convencionais de perfilagem de poço. Com a aplicação de técnicas de sobreamostragem e validação, os métodos baseados em ensemble apresentaram desempenho superior, destacando-se na identificação do folhelho silicoso orgânico, litofácies de maior interesse petrolífero na formação analisada.

Já Narayan et al. (2023) aplicaram cinco técnicas a dados sísmicos 3D para classificar litofácies. No campo Penobscot, o MLP obteve o melhor desempenho geral, enquanto o AdaBoost apresentou os piores resultados. As classificações obtidas demonstraram boa concordância com dados de poço, evidenciando a eficácia dessas abordagens.

Cabral et al. (2025) propuseram o uso de aprendizado de máquina com algoritmo genético e busca exaustiva para otimização de modelos aplicados à classificação litológica e predição de TOC, utilizando dados de um poço do Campo de Marlim. O XGB obteve o melhor desempenho em ambas as tarefas, destacando o potencial dessas abordagens na caracterização de reservatórios.

Diante dos estudos apresentados, observa-se que os métodos de aprendizado de máquina têm se destacado na tarefa de classificação litológica, especialmente quando combinados com técnicas de pré-processamento. Nesse contexto, este trabalho tem como objetivo avaliar o uso de algoritmos de aprendizado de máquina na classificação de litologias, a partir de dados de três poços do Campo de Marlim, utilizando o *K-Means Featurizer* para o agrupamento e a geração de novos atributos.

O artigo está organizado da seguinte forma: a Seção 2 apresenta os materiais e métodos, detalhando os dados utilizados e a metodologia empregada; a Seção 3 discute os resultados obtidos; e a Seção 4 apresenta as conclusões finais.

2. MATERIAIS E MÉTODOS

2.1 Descrição e origem dos dados utilizados

Os dados utilizados neste trabalho foram obtidos de registros de poços do Campo de Marlim, na porção nordeste da Bacia de Campos, litoral do Rio de Janeiro. A área de estudo cobre cerca de 257,6 km^2 e as informações foram fornecidas pela Agência Nacional do Petróleo (ANP).

Foram analisados dados de três poços, totalizando 5.869 amostras, cada uma contendo oito atributos petrofísicos. Os atributos utilizados no estudo são: GR (Perfil de Raio Gama), NPHI (Perfil Neutrônico), RHOB (Perfil de Densidade), DRHO (Delta Densidade), ILD (Perfil de Indução), SFLU (Resistividade), CALI (Calibre do Poço) e DT (Perfil Sônico). Para a análise, foram selecionadas as classes litológicas comuns aos três poços: argilito, marga e arenito.

A Tabela 1 apresenta a contagem das amostras por classe geológica em cada poço, bem como o total geral.

Tabela 1: Contagem de amostras por classe em cada poço e total geral

Classe	**Poço 1**	**Poço 2**	**Poço 3**	**Total Geral**
Argilito	327	958	1393	2678
Marga	1369	471	778	2618
Arenito	263	112	198	573
Total de amostras	**1959**	**1541**	**2369**	**5869**

Fonte: Elaborado pelos autores (2025).

2.2 Pré-processamento

No pré-processamento dos dados, foi adotado o *K-Means Featurizer*, proposto por Kouadio et al. (2024). Essa técnica aplica o algoritmo de clusterização *K-Means* para gerar novas representações de características, combinando os atributos de entrada com a variável alvo. A abordagem amplia o *K-Means* tradicional ao incorporar informações supervisionadas durante o agrupamento, por meio da concatenação da variável alvo aos atributos originais. Esse procedimento resulta em um espaço aumentado, no qual cada amostra é representada não apenas pelos atributos originais, mas também pelas distâncias em relação aos centróides dos clusters formados.

Para garantir que variáveis com escalas distintas não influenciassem na formação dos agrupamentos, os dados foram normalizados exclusivamente para a etapa de clusterização. As novas características geradas foram, então, incorporadas à base original, sendo utilizadas como atributos complementares nas etapas posteriores de classificação.

2.3 Divisão dos Dados

A base de dados foi dividida em conjuntos de treinamento e teste. O conjunto de treinamento é usado para o aprendizado dos modelos, enquanto o conjunto de teste serve para avaliar seu desempenho em dados não vistos, verificando sua capacidade de generalização. Após a inclusão dessa nova característica na base (por meio do *K-Means Featurizer*), a amostra foi dividida nas proporções de 80% para treinamento e 20% para teste. Essa divisão foi realizada de forma aleatória e estratificada, preservando a proporção original das classes litológicas em ambos os conjuntos.

2.4 Métodos de Classificação Supervisionada

Seis algoritmos supervisionados de Aprendizado de Máquina foram utilizados para a classificação dos dados: *K-Nearest Neighbors (KNN)*, *Support Vector Machines (SVM)*, *Decision Tree (DT)*, *MultiLayer Perceptron (MLP)*, *Extreme Gradient Boosting (XGB)* e *Natural Gradient Boosting (NGB)*. Cada um apresenta características específicas quanto à capacidade de modelagem, complexidade e abordagem de aprendizado.

O *Decision Tree* é um algoritmo supervisionado, não paramétrico. O modelo constrói uma estrutura hierárquica onde cada nó interno realiza divisões baseadas em características dos dados, e as folhas indicam as previsões finais. Essa abordagem divide recursivamente o espaço de entrada em regiões distintas, facilitando a interpretação por meio de regras lógicas do tipo "se-então" (Semanjski, 2023).

O algoritmo KNN é um método supervisionado que classifica uma nova amostra com base na proximidade dos seus k vizinhos mais próximos no conjunto de treinamento. Utilizando geralmente a distância Euclidiana para medir essa proximidade, o KNN atribui à amostra o rótulo mais frequente entre os vizinhos. O valor de k influencia diretamente o desempenho do algoritmo, com valores muito baixos ou muito altos podendo afetar a precisão (Qaisar, 2023).

O MLP é uma rede neural artificial do tipo *feedforward*, composta por camadas de neurônios que aplicam funções de ativação não lineares, como a sigmoide ou ReLU. Essa arquitetura permite modelar relações complexas e não lineares entre as variáveis de entrada e saída. O treinamento do MLP é realizado por meio do algoritmo de retropropagação do erro, que ajusta os pesos da rede minimizando uma função de custo associada à diferença entre as saídas previstas e os valores reais. Esse processo iterativo permite que o modelo aprenda a mapear padrões complexos nos dados para realizar tarefas de classificação (Chan et al., 2023).

O NGB é uma variante do *gradient boosting* que utiliza o gradiente natural, uma generalização do gradiente tradicional que considera a geometria da superfície paramétrica para otimizar a função de perda. Essa abordagem é utilizada em modelos probabilísticos, permitindo a estimação da incerteza nas previsões e melhor convergência (Duan et al., 2020).

O gradiente natural é obtido multiplicando o gradiente tradicional pela inversa da matriz de informação de Fisher, conforme a equação 1:

$$\tilde{\nabla}\mathcal{L}(\theta, y) \propto \mathcal{I}_{\mathcal{L}}(\theta)^{-1}\nabla\mathcal{L}(\theta, y), \tag{1}$$

onde $\mathcal{I}_{\mathcal{L}}(\theta)$ representa a matriz de informação de Fisher associada à função de perda $\mathcal{L}$.

O SVM busca encontrar o hiperplano ótimo que separa as classes com a maior margem possível, definido pela equação 2:

$$f(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + b, \tag{2}$$

onde $\mathbf{w}$ é o vetor normal ao hiperplano e b é o deslocamento.

A margem máxima é obtida resolvendo-se um problema de otimização convexa. O uso de funções núcleo *(kernels)* permite lidar com separações não lineares ao mapear os dados para espaços de dimensão superior, favorecendo a separação entre as classes (Otchere et al., 2021).

O XGB é um algoritmo de boosting baseado em árvores de decisão que combina modelos fracos sequencialmente para minimizar uma função de perda. Seu diferencial está na inclusão de um termo de regularização para controlar a complexidade do modelo e evitar sobreajuste, além do uso da segunda derivada da função de perda, o que permite uma otimização mais precisa pelo método de Newton (Chen and Guestrin, 2016).

A função objetivo do XGB, apresentada na Equação 3:

$$\mathcal{L}(\phi) = \sum_i l(\hat{y}i, y_i) + \sum k\Omega(f_k), \tag{3}$$

onde o termo de regularização $\Omega(f_k) = \gamma T + \frac{1}{2}\lambda|w_k|^2$ penaliza a complexidade do modelo, sendo T o número de folhas da árvore, w_k os pesos das folhas, γ o custo por folha e λ o parâmetro de regularização.

2.5 Validação Cruzada

A validação cruzada *k-fold* é uma técnica de reamostragem utilizada para estimar o erro de generalização de um modelo preditivo. O procedimento consiste em dividir o conjunto de dados em k subconjuntos de tamanho aproximadamente igual. Em cada uma das k iterações, o modelo é ajustado com k – 1 subconjuntos (conjunto de treinamento) e avaliado no subconjunto restante (conjunto de teste). A estimativa final do erro de predição é obtida pela média dos erros calculados em cada partição (Hastie et al., 2009)

2.6 Otimização de Hiperparâmetros com Grid Search

O *Grid Search* é uma técnica que realiza a avaliação exaustiva de todas as combinações possíveis de hiperparâmetros dentro de uma grade predefinida, com o objetivo de identificar a configuração que maximiza o desempenho do modelo preditivo (Bergstra and Bengio, 2012). O método permite a exploração do espaço definido de parâmetros, garantindo que nenhuma combinação seja ignorada.

Cada experimento foi realizado 30 vezes, de forma independente, utilizando diferentes sementes aleatórias para a divisão dos dados. Além disso, foi empregada a validação cruzada do tipo *k-fold* com $k = 5$. Os hiperparâmetros avaliados estão descritos na Tabela 2.

Tabela 2: Hiperparâmetros dos modelos de classificação

Método	**Parâmetros**	**Configuração**
DT	criterion	[gini, entropy]
	max_depth	[3, 5, 10, None]
KNN	n_neighbors	[1, 2, 3, 4, 5, 6, 10]
	weights	[uniform, distance]
MLP	hidden_layer_sizes	[(50,), (50, 50), (50, 50, 50), (50, 50, 50, 50), (100,), (100, 100), (100, 100, 100), (100, 100, 100, 100)]
	activation	[identity, logistic, tanh, relu]
NGB	n_estimators	[50, 100, 200]
	base	b1 = DecisionTreeRegressor (criterion="friedman_mse", max_depth=2) b2 = DecisionTreeRegressor (criterion="friedman_mse", max_depth=4)
SVM	kernel	[linear, rbf]
	gamma	[0,01; 0,1; 1]
	C	[1, 10, 100]
XGB	n_estimators	[50, 100, 200]
	learning_rate	[0,01; 0,1; 0,2]
	max_depth	[3, 5, 10]

Fonte: Elaborado pelos autores (2025).

3. Métricas de Avaliação

Para avaliar o desempenho dos modelos, foram utilizadas as métricas de acurácia, precisão, *recall* e *F1-score*. A Tabela 3 apresenta uma breve descrição de cada uma. Essas métricas permitem analisar aspectos variados do desempenho, como a capacidade geral de acerto e o equilíbrio na identificação correta das classes, especialmente em problemas com desbalanceamento.

Tabela 3: Descrição das métricas de avaliação utilizadas

Métrica	**Descrição**
Acurácia	Proporção de previsões corretas em relação ao total de amostras avaliadas.
Precisão	Proporção de verdadeiros positivos entre todas as amostras classificadas como positivas pelo modelo.
Recall	Proporção de verdadeiros positivos em relação ao total de amostras que realmente são positivas.
F1-score	Média harmônica entre a precisão e o *recall*, equilibrando ambos os aspectos.

Fonte: Elaborado pelos autores (2025).

4. RESULTADOS E DISCUSSÃO

A Tabela 4 apresenta a distribuição da coluna clusters adicionada no conjunto de treino após a aplicação do *k-Means Featurizer*.

Tabela 4: Distribuição dos *clusters* após normalização e aplicação do *k-Means Featurizer* no conjunto de treinamento.

Cluster	**Quantidade de Amostras**
0	1.391
1	1.591
2	1.713
Total	**4.695**

Fonte: Elaborado pelos autores (2025).

Para a avaliação do desempenho dos modelos de classificação litológica, foram utilizadas as métricas de Acurácia, Precisão, Recall e F1-Score. As Tabelas 5 e 6 apresentam os resultados médios obtidos nos conjuntos de treinamento e teste, respectivamente.

Observa-se que, no conjunto de treinamento, o algoritmo XGB obteve o melhor desempenho, com acurácia = 0,893, precisão = 0,894, *recall* = 0,893 e *F1-Score* = 0,893. No conjunto de teste, o SVM teve os piores resultados e foi o único que apresentou queda em relação ao desempenho no treinamento. Já o NGB, que havia sido o pior no treinamento, apresentou melhora no teste e superou o SVM em todas as métricas. O XGB também apresentou os melhores resultados no teste, atingindo acurácia = 0,910, precisão = 0,910, *recall* = 0,910 e *F1-Score* = 0,910, mostrando sua superioridade e capacidade de generalização em relação

aos demais métodos avaliados. Essa superioridade foi estatisticamente comprovada pelo teste de Wilcoxon pareado, que indicou diferenças significativas, com *p*-valor inferior a 0,05, em relação aos demais modelos.

Tabela 5: Média e desvio padrão das métricas - conjunto de treinamento. Os melhores resultados estão destacados em negrito, enquanto o símbolo * indica que a diferença em relação ao melhor resultado é estatisticamente significativa. Considera-se que dois conjuntos de resultados são estatisticamente diferentes quando o p-valor do teste de Wilcoxon é inferior a 0,05.

Modelo	**Acurácia**	**Precisão**	**Recall**	**F1-Score**
DT	0,835 (0,010)*	0,836 (0,010)*	0,835 (0,010)*	0,835 (0,010)*
KNN	0,825 (0,012)*	0,827 (0,012)*	0,825 (0,012)*	0,825 (0,012)*
MLP	0,807 (0,012)*	0,810 (0,012)*	0,807 (0,012)*	0,807 (0,012)*
NGB	0,750 (0,014)*	0,758 (0,013)*	0,750 (0,014)*	0,750 (0,013)*
SVM	0,795 (0,009)*	0,799 (0,009)*	0,795 (0,009)*	0,796 (0,009)*
XGB	**0,893** (0,008)	**0,894** (0,008)	**0,893** (0,008)	**0,893** (0,008)

Fonte: Elaborado pelos autores (2025).

Tabela 6: Melhores Modelos - Conjunto de Teste

Modelo	**Melhores hiperparâmetros**	**Acurácia**	**Precisão**	**Recall**	**F1-Score**
DT	criterion='gini', max_depth=10	0,836	0,836	0,836	0,836
KNN	n_neighbors=4, weights='distance'	0,836	0,836	0,836	0,836
MLP	activation='relu', hidden_layer_sizes=(100, 100, 100)	0,824	0,824	0,824	0,824
NGB	Base=DecisionTreeRegressor (criterion='friedman_mse', max_depth=2), n_estimators=50	0,769	0,773	0,769	0,769
SVM	C=100, gamma=1, kernel='rbf'	0,731	0,745	0,731	0,728
XGB	learning_rate=0,2, max_depth=10, n_estimators=200	0,910	0,910	0,910	0,910

Fonte: Elaborado pelos autores (2025).

A aplicação do *K-Means Featurizer* possibilitou a adição de uma nova variável aos dados: o identificador do *cluster* atribuído a cada amostra. Contudo, a avaliação da importância das variáveis por meio dos valores SHAP, ilustrada na Figura 1, indica que a variável de *cluster* não teve a relevância esperada na maioria dos modelos.

Apesar disso, os resultados mostram que o *K-Means Featurizer* conseguiu capturar certas informações úteis, embora seu impacto na performance geral dos modelos tenha sido limitado. Ainda assim, a variável *cluster* figurou entre os atributos relevantes, sugerindo que sua inclusão contribuiu, ainda que de forma modesta, para a modelagem.

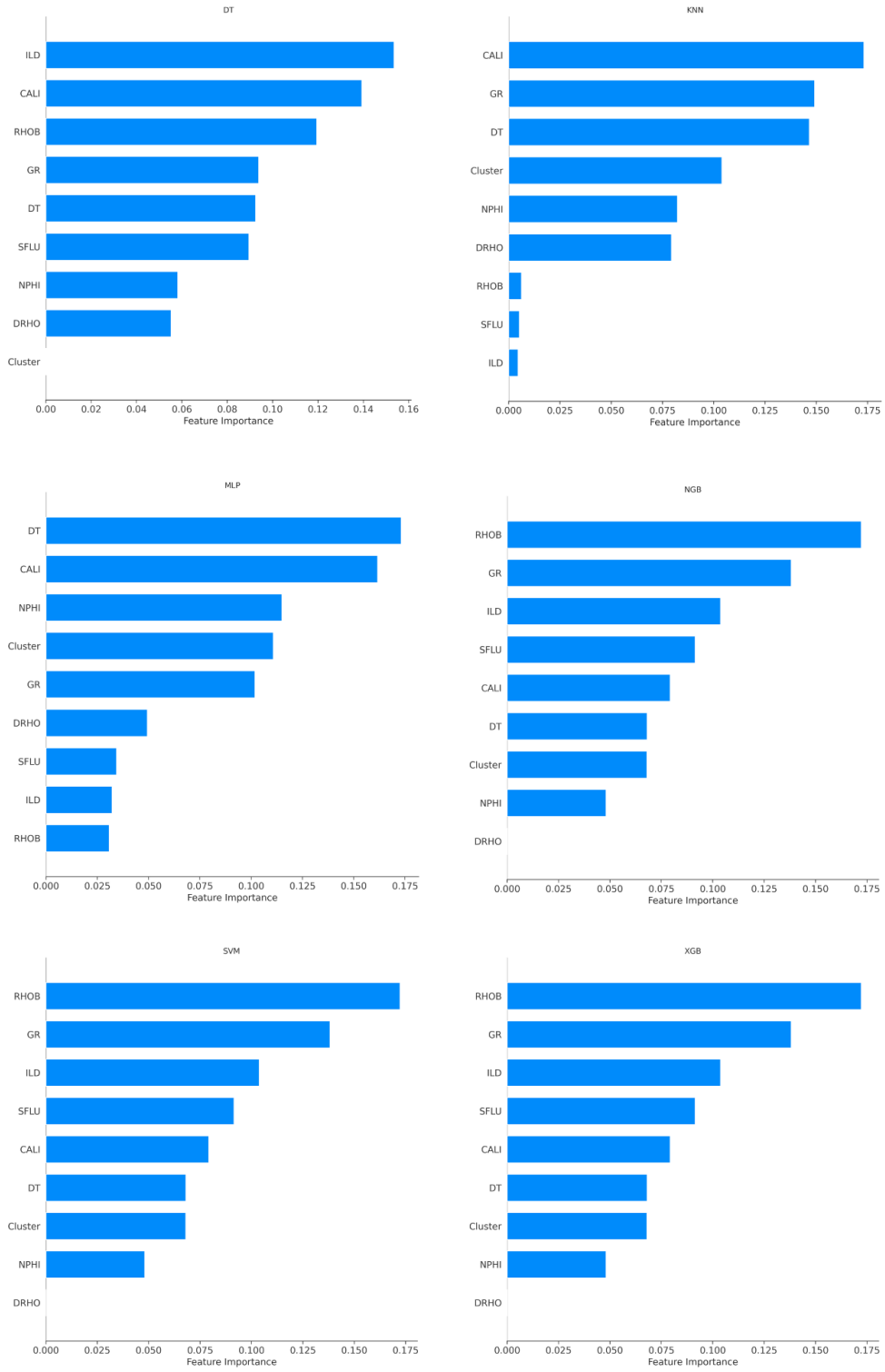

Figura 1: Importância das variáveis calculada por análise SHAP para os modelos de aprendizado supervisionado. Cada barra representa o impacto médio de cada atributo, indicando a contribuição relativa de cada atributo para o desempenho do modelo.

5. CONCLUSÃO

Este estudo avaliou o uso da técnica de extração de características *K-Means Featurizer*, combinada com o *Grid Search*, aplicada a seis algoritmos de aprendizado de máquina para um problema de classificação litológica, utilizando dados coletados no Campo de Marlim, na Bacia de Campos. Após o processo de treinamento e validação, foram calculadas as métricas médias de desempenho para comparar a eficácia dos modelos. O algoritmo que apresentou os melhores resultados, tanto nos conjuntos de treino quanto de teste, foi o XGB.

Com base nos resultados obtidos, o modelo treinado pode ser aplicado a outros conjuntos de dados semelhantes para automatizar a interpretação litológica, reduzindo erros e a subjetividade da análise manual. Isso torna a classificação dos perfis de poços mais rápida, acelerando a avaliação do reservatório e, consequentemente, antecipando o início da produção de petróleo. Além disso, o uso desse modelo pode contribuir para a redução de custos operacionais, ao diminuir retrabalhos e permitir análises mais confiáveis.

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 também agradecem à Fundação Carlos Chagas Filho de Amparo à Pesquisa do Estado do Rio de Janeiro (FAPERJ), bem como à Agência Nacional do Petróleo, Gás Natural e Biocombustíveis (ANP).

Referências

Abbas, Z. and Rasool, M. (2024). Petrophysical properties in reservoir evaluation: Insights from well-log analysis and data interpretation for enhanced hydrocarbon assessment. *ResearchGate Preprint*.

Ambagtsheer, R. C., Shafiabady, N., Dent, E., Seiboth, C., and Beilby, J. (2020). The application of artificial intelligence (ai) techniques to identify frailty within a residential aged care administrative data set. *International Journal of Medical Informatics*, 136.

Bergstra, J. and Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, 13:281–305.

Cabral, J. d. C., Silva, C. A. d., de Souza, G., and Martins Saporetti, C. (2025). Uso de métodos de aprendizado de máquina e algoritmo genético para predição de toc e classificação de litologia. *VETOR - Revista De Ciências Exatas E Engenharias*, 35(1):e18357.

Chan, K. Y., Abu-Salih, B., Qaddoura, R., Al-Zoubi, A. M., Palade, V., Pham, D.-S., Ser, J. D., and Muhammad, K. (2023). Deep neural networks in the cloud: Review, applications, challenges and research directions. *Neurocomputing*, 545.

Chen, G., Lu, C., Wang, Q., Du, G., and Chen, J. (2010). Characteristics of pore evolution and its controlling factors of baiyun sag in deepwater area of pearl river mouth basin. *Acta Petrologica Sinica*, 31:566–572.

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. *CoRR*, abs/1603.02754.

Duan, Y., Xie, J., Su, Y., Liang, H., Hu, X., Wang, Q., and Pan, Z. (2020). Application of the decision tree method to lithology identification of volcanic rocks—taking the mesozoic in the laizhouwan sag as an example. *Scientific Reports*, 10(1).

Fu, G., Yan, J., Zhang, K., Hu, H., and Luo, F. (2017). Current status and progress of lithology identification technology. *Progress in Geophysics*, 32:26–40.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2nd edition.

Hou, M., Xiao, Y., Lei, Z., Yang, Z., Lou, Y., and Liu, Y. (2023). Machine learning algorithms for lithofacies classification of the gulong shale from the songliao basin, china. *Energies*, 16(6).

Hu, Q., Wang, L., and Cui, E. (2013). Tight reservoir evaluation of the unconventional oil and study on the main controlling factors of hydrocarbon accumulation in santanghu basin. *Petroleum Geology and Oilfield Development in Daqing*, 32:164–169.

Ibrahim, A., Gamble, P., Jaroensri, R., Abdelsamea, M. M., Mermel, C. H., Chen, P.-H. C., and Rakha, E. A. (2020). Artificial intelligence in digital breast pathology: techniques and applications. *The Breast*, 49:267–273.

Kouadio, L., Liu, J., Liu, R., Wang, Y., and Liu, W. (2024). K-means featurizer: A booster for intricate datasets. *Earth Science Informatics*, 17:1–26.

Narayan, S., Konka, S., Chandra, A., Abdelrahman, K., Andráš, P., and Eldosouky, A. M. (2023). Accuracy assessment of various supervised machine learning algorithms in litho-facies classification from seismic data in the penobscot field, scotian basin. *Frontiers in Earth Science*, Volume 11 - 2023.

Otchere, D. A., Arbi Ganat, T. O., Gholami, R., and Ridha, S. (2021). 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*, 200.

Qaisar, S. M. (2023). 13 - adaptive rate eeg processing and machine learning-based efficient recognition of epilepsy. In Pal, K., Ari, S., Bit, A., and Bhattacharyya, S., editors, *Advanced Methods in Biomedical Signal Processing and Analysis*, pages 341–373. Academic Press.

Saporetti, C. M., Goliatt, L., and Pereira, E. (2021). Neural network boosted with differential evolution for lithology identification based on well logs information. *Earth Science Informatics*, 14(1):133–140.

Semanjski, I. C. (2023). Chapter 5 - data analytics. In Semanjski, I. C., editor, *Smart Urban Mobility*, pages 121–170. Elsevier.

Xia, M., Jiang, C., Qian, Z., Xia, Z., Wang, B., and Sun, T. (2010). Geochemistry and petrogenesis of huangshandong intrusion, east tianshan, xinjiang. *Acta Petrologica Sinica*, 26:2413–2430.

Application of Machine Learning Techniques for Lithology Classification of Oil Wells

Abstract.*Lithological classification is a fundamental step in the characterization of petroleum reservoirs, as it allows for the identification of rock types and their physical and mineralogical properties, contributing to the analysis of storage capacity and hydrocarbon flow. Traditional methods, such as the manual interpretation of geophysical logs, are still used but face limitations due to their reliance on human interpretation and lack of efficiency when handling large volumes of data. In this context, the use of machine learning techniques has gained prominence by enabling the automated and accurate detection of complex patterns. This study proposes a computational approach for lithological classification using supervised learning algorithms combined with the K-Means Featurizer feature extraction technique. Data from three wells in the Marlim Field, located in the Campos Basin, were used. The models were optimized with Grid Search and validated using K-fold cross-validation. Model performance was evaluated based on accuracy, precision, recall, and F1-score metrics. The XGB algorithm achieved the best performance, reaching an accuracy of 0.910. The results highlight the potential of combining clustering methods with supervised models to improve lithological classification in sedimentary contexts with challenging interpretation.*

Keywords: *Lithology Classification, Machine Learning, K-Means Featurizer, Oil Reservoirs*