INTELIGÊNCIA ARTIFICIAL APLICADA À PREDIÇÃO DE FÁCIES EVAPORÍTICAS NA BACIA DE SANTOS: UMA ANÁLISE DE ÁRVORES DE CLASSIFICAÇÃO

José Augusto Vitorino Dias[1], Mário Martins Ramos[1], Antonio Fernando Menezes Freire[1,2], Wagner Moreira Lupinacci[1,2]

[1] Grupo de Interpretação Exploratória e Caracterização de Reservatórios (GIECAR) da Universidade Federal Fluminense (UFF)

[2] Instituto Nacional de Ciência e Tecnologia de Geofísica do Petróleo (INCT-GP/CNPQ)

Resumo

Fácies evaporítica são difíceis de perfurar devido ao comportamento mecânico dos seus constituintes minerais. Portanto, uma classificação precisa de evaporitos pode representar uma grande vantagem para a atividade de perfuração de poços. O objetivo deste trabalho consiste em classificar fácies evaporíticas da Formação Ariri, da Bacia de Santos, utilizando métodos de inteligência artificial derivados de classificadores de árvores. Os métodos escolhidos foram: Árvore de Decisão, Random Forest e *Catboost*. Os algoritmos foram treinados para identificar quatro fácies evaporítica: halita, anidrita, carnalita e taquidrita e para isso nós utilizamos os perfis de raio gama (GR), velocidade compressional (VP) e velocidade cisalhante (VS) como os atributos de entrada. A avaliação foi feita partindo de uma sessão já categorizada para comparação de acertos e erros utilizando as métricas de acurácia e F1-*score*. Todos os métodos apresentaram resultados com média acima de 85% em ambas métricas. A partir de análises das métricas, observamos que o método CatBoost obteve o melhor resultado.

Summary

Evaporitic facies are difficult to drill due to the mechanical behavior of their mineral constituents. Therefore, a precise classification of evaporites can represent a great advantage for drilling activities. The objective of this work is to classify the evaporitic facies of the Ariri Formation in the Santos Basin, using machine learning methods derived from tree classifiers. The chosen methods were: Decision Tree, Random Forest, and Catboost. These algorithms were trained to identify four evaporitic facies: Halite, Anhydrite, Carnalite, and Tachydrite, and for that we use the well logs: gamma-ray (GR), compressional velocity (VP), and shear (VS) velocity. The evaluation was conducted by comparing the results with an already categorized dataset, using accuracy and F1-score metrics. All methods had averages above 85% in both metrics. From metrics analysis, we observed that the CatBoost method obtained the best result.

Introdução

A classificação de fácies é uma tarefa complexa que demanda tempo e conhecimentos específicos. É possível adquirir as propriedades físicas das rochas por meio de métodos indiretos, mesmo assim faz-se necessário saber quais são as fácies litológicas presentes no poço a ser estudado. Nesse sentido, o uso de *machine learning* permite a obtenção de fácies a partir da utilização dos dados de perfis geofísicos. Inserido nos métodos de *machine learning* existe uma categoria baseada em árvores de decisões. Esses métodos têm sido bastante utilizados por apresentar bons resultados e ter custo computacional reduzido (Ho, 1995; Vikrant e Eden, 2019; Silva, 2022).

A árvore de decisão, segundo Quinlan (1986), é um método de *machine learning* menos complexo que os demais, e capaz de resolver problemas de alta complexidade. Para Lantz (2015), a árvore de decisões é capaz de resolver problemas de predileção e ao mesmo tempo simples de serem compreendidas, visto que podem ser demonstradas visualmente, sem a necessidade de muito conhecimento matemático.

O método de árvore de decisão começou a ser desenvolvido na década de 1960 por Hunt–Szymanski pensado em modelar a aprendizagem humana. Esse método foi utilizado como modelo para os algoritmos baseados em árvore de decisão mais avançados, como por exemplo:

Discotomiser Iterative 3 (ID3), criado por Quinlan (1986), divide as características dos previsores do dado utilizando a entropia para criar os melhores nós, um componente da árvore de decisão. As informações do dado são fornecidas pelo *Information Gain*, selecionando o maior ganho de informações para a melhor classificação.

O C4.5, também desenvolvido por Quinlan (1994), é uma extensão do ID3. O C4.5 funciona selecionando a melhor característica para dividir os dados em cada nó da árvore, através do conceito de entropia. Diferentemente do ID3, o C4.5 consegue classificar dados categóricos e numéricos, além de conseguir processar dados ausentes.

Classification and Regression Trees (CART) é um método de classificação e regressão baseado em árvore de decisões. Implementado por Breiman (1984), esse modelo usa um processo iterativo para dividir o conjunto de dados em subconjuntos menores e constrói a árvore de decisão.

A Floresta Aleatória é uma classificação que concatena várias árvores aleatórias para obter a classificação e o resultado é uma média de todas as árvores. Teorizado por Ho (1995), a Floresta Aleatória é uma extensão da agregação *bootstrap*, um método de re-seleção randômico, pois a aleatoriedade com que os dados são re-selecionados faz com que não ocorra correlação entre as árvores aleatórias, melhorando o seu resultado.

Os algoritmos de *Boosting* são classificadores que começaram a ser desenvolvidos por Leo Breiman. Estes algoritmos enfatizam o treinamento em amostras que foram erroneamente classificadas, atribuindo uma ponderação diretamente associada ao erro (Sammut e Webb, 2010). Dentro dessa família existem os classificadores AdaBoost, CatBoost, XGBoost.

Na geociência, o uso de técnicas de aprendizado de máquina tem se tornado amplamente difundido devido à sua capacidade de agilizar processos e aprimorar métricas já existentes. Um exemplo relevante é o estudo realizado por Silva (2022) na classificação de eletrofácies, no qual foi empregada a técnica de floresta aleatória. Nesse estudo, as eletrofácies de brecha, conglomerado, quartzito, xisto e ultramáficas foram classificadas utilizando parâmetros de entrada como densidade, susceptibilidade magnética e condutividade elétrica.
Os resultados obtidos foram bastante promissores, evidenciando uma acurácia de 0,86 na classificação das eletrofácies. Essa alta taxa de acerto demonstra a eficácia da abordagem de aprendizado de máquina aplicada no estudo, proporcionando uma ferramenta para a interpretação de dados geocientíficos. Ao adotar técnicas como a floresta aleatória, é possível extrair informações confiáveis sobre a composição e características das eletrofácies estudadas, contribuindo assim para um melhor entendimento dos processos geológicos envolvidos.

Vikrant e Eden (2019) compararam a performance de três métodos baseados em árvore (XGBoost, LightGBM e CatBoost) para a classificação de oito fácies: rocha carbonática, carvão, arenito seixoso, conglomerado, arenito médio, arenito fino, siltito e lamito na Bacia de Ordos (China). Eles utilizaram os perfis de gamma ray, sônico, densidade, neutrão, laterolog profundo, laterolog raso e cáliper. Foi aplicada uma otimização de hiperparâmetros, e a avaliação dos resultados foi feita através da matriz de confusão, precisão, revocação e F1-score. O LightGBM foi o método que apresentou a melhor performance. Além disso, o resultado indicou que o CatBoost é capaz de lidar com grandes conjuntos de dados de forma eficiente, sem sacrificar o desempenho ou a capacidade de treinar modelos complexos em um tempo razoável.

Os resultados anteriormente mencionados ressaltam a eficiência dos métodos baseados em árvores de decisão. Diante desse contexto, o presente estudo tem como objetivo principal comparar diferentes métodos baseados em árvores de decisão para a classificação de fácies evaporíticas na Formação Ariri da Bacia de Santos. Os métodos avaliados nesta pesquisa incluem Árvore de Decisão, Floresta Aleatória e CatBoost.

Metodologia

Antes de performar a classificação, nós realizamos um pré-processamento dos dados definindo o alvo (fácies) e os perfis geofísicos previsores (raios gama - GR, velocidade compressional - VP e velocidade cisalhante - VS). Nessa etapa, os previsores foram escalonados para média zero e desvio padrão 1, ou seja, para que eles possam ficar em uma escala comum. O conjunto de dados foi dividido em duas partes: uma parte para ajustar o modelo de previsão e a outra para avaliar o modelo construído. Para isso, utilizamos 30% dos dados para treinar o algoritmo e 70% dos dados para classificar as fácies. Além dessa divisão, foi utilizada uma métrica de validação cruzada que redimensiona o dado mantendo a proporção de fácies. Essa técnica serve para avaliar o desempenho dos métodos. Em relação ao custo computacional, a Árvore de Decisão tem menor custo.

Os resultados foram avaliados utilizando a acurácia, a matriz confusão e o F1-score. A acurácia indica para o usuário qual modelo teve o maior percentual de acertos, sendo que seu cálculo é aplicado para todas as fácies. Já o *F1-Score* é a média harmônica de outros dois parâmetros, a precisão e a revocação. O *F1-Score* é calculado para cada fácies.

Os algoritmos foram implementados na linguagem de programação Python, utilizando uma variedade de bibliotecas para facilitar o desenvolvimento e a análise de dados. As principais bibliotecas utilizadas foram o NumPy, para manipulação eficiente de *arrays* e cálculos numéricos, o Matplotlib, para visualização de gráficos e plotagem de resultados, e o Scikit-learn, uma biblioteca amplamente reconhecida para aprendizado de máquina e mineração de dados. A combinação dessas bibliotecas forneceu um ambiente robusto e eficiente para a implementação dos algoritmos, permitindo uma análise completa e visualização adequada dos resultados obtidos.

Resultados

O dataset em questão apresenta como característica principal a presença predominante de Halita, representando cerca de 84,34% das ocorrências. Em sequência, temos a Anidrita com 7,13%, a Carnalita com 6,75% e a Taquidrita com 1,78%. Essas porcentagens refletem a distribuição das diferentes espécies de sal presentes no conjunto de dados.

Analisando os gráficos de distribuição dos previsores (atributos de entrada - Figura 1), temos que os maiores valores de GR (com média em torno de 71 API) são referentes à Carnalita, provavelmente, devido a presença do isótopo radioativo K40 em sua composição (Figura 1-a). Já os outros sais apresentam valores de GR mais baixos, em torno de 15 API. Nas velocidades cisalhante

(VS) e compressional (VP), apresentadas na Figuras 1 b) e c), podemos observar que os sais que possuem baixas velocidades são a Taquidrita (médias de 3600 m/s e 1800 m/s para VP e VS, respectivamente) e Carnalita (médias de 4100 m/s e 2200 m/s para VP e VS, respectivamente). A distribuição da Halita mostra valores intermediários, com velocidades aproximadas de 4500 e 2500 m/s e baixo desvio padrão (112 m/s e 72 m/s para VP e VS, respectivamente). As velocidades mais altas são associadas a Anidrita, cujos valores médios são de 5300 m/s e 2800 m/s para VP e VS, conforme apresentado na Figura 1 b) e Figura 1 c) respectivamente.

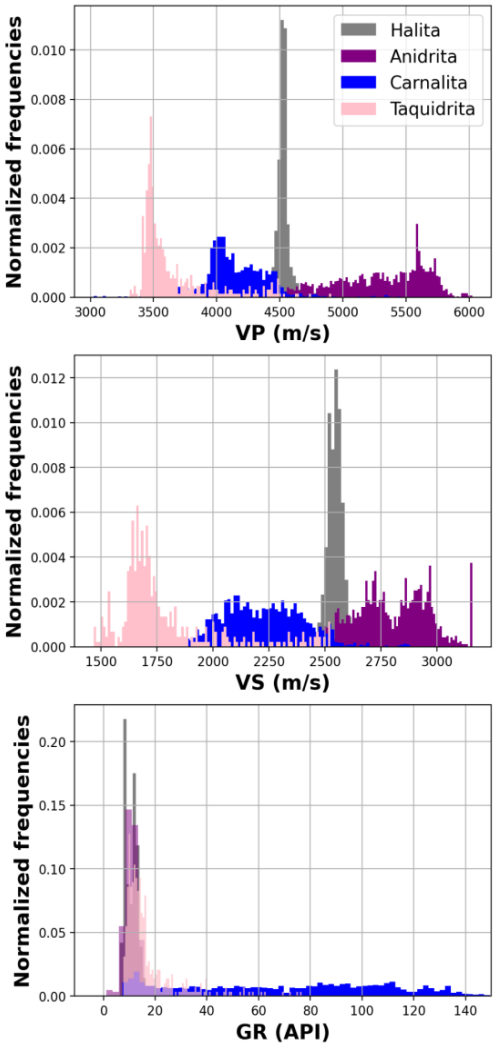

Figura 1 *- Distribuição relativa normalizada de frequência dos previsores utilizados: (a) GR, (b) VP e (c) VS.*

Ao analisarmos as métricas, observamos que o F1-Score (Figura 2) revela um desempenho superior na identificação da fácie de Halita. Essa superioridade pode ser atribuída, em parte, à maior quantidade de amostras disponíveis no conjunto de dados para essa fácies específica. Por outro lado, a fácies de taquidrita, com a menor quantidade de amostras, obteve o pior desempenho na sua estimativa. Essa discrepância nos resultados pode estar associada a uma possível influência do *F1-Score*, que tende a levar em consideração a quantidade de dados disponíveis. É importante destacar que o *F1-Score* é uma métrica que combina a precisão e a taxa de recuperação, fornecendo uma medida balanceada do desempenho do modelo. No entanto, quando a quantidade de amostras é desigual entre as classes, o *F1-Score* pode ser afetado, resultando em estimativas menos precisas para as classes com menor representatividade.

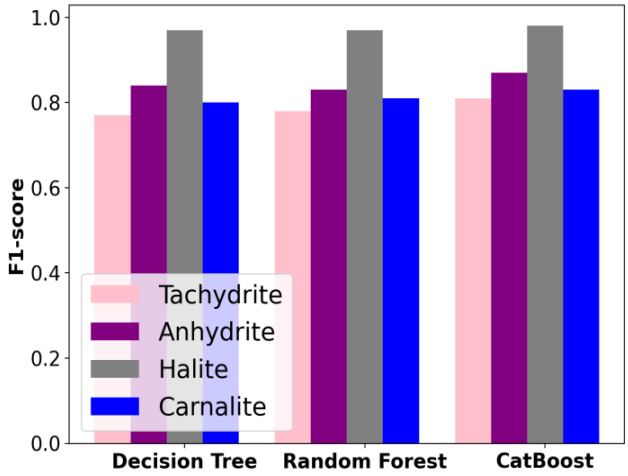

Figura 2 - *Resultados do valor-f1 por fácies para cada método.*

Ao observar a métrica de acurácia (Figura 3), podemos notar que os três métodos apresentam valores bastante semelhantes. Isso indica que o número de amostras das fácies não exerceu uma influência significativa nessa métrica de desempenho. Todos os métodos alcançaram uma acurácia acima de 80% na classificação das fácies. Esses resultados são encorajadores, pois indicam uma consistência geral no desempenho dos métodos em identificar corretamente as fácies. A acurácia, como métrica, fornece uma medida geral da taxa de acertos do modelo em relação ao total de amostras. O fato de todos os métodos terem atingido acurácias superiores a 80% é um indicativo positivo de sua eficácia na classificação das fácies.

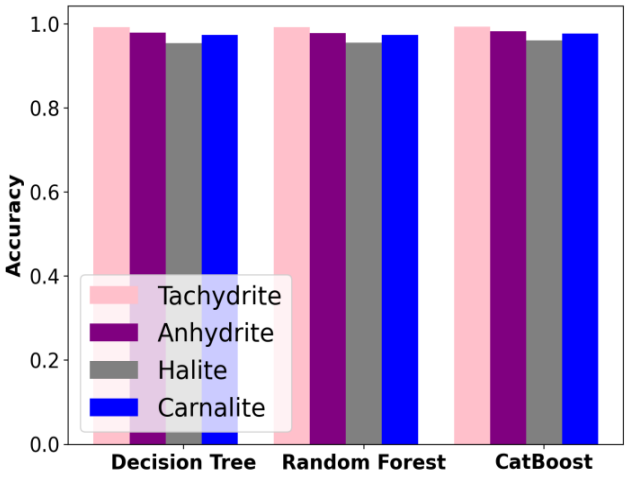

Figura 3 - *Resultados da acurácia por fácies para cada método.*

No entanto, é importante ressaltar que a acurácia por si só pode não fornecer informações completas sobre o desempenho do modelo, especialmente quando há classes desbalanceadas ou características específicas de cada fácies que podem impactar a interpretação geológica.

Ao analisar a Figura 4, podemos observar que os métodos utilizados apresentaram resultados com classificações próximas ao gabarito (perfil de fácies verdadeiras). Além disso, notamos que o número de amostras classificadas para cada fácies foi muito próximo à quantidade de amostras presentes no gabarito.

É importante destacar que a maior variação no número de amostras ocorreu na classificação da fácie anidrita, especialmente nas profundidades em torno de 3600 m e 3850 m (Figura 5). Isso indica que os métodos empregados tiveram mais dificuldade em classificar corretamente essa fácies, o que pode ser atribuído que a anidrita dificilmente ocorre na sua forma “pura”, muitas vez, ela é uma mistura anidrita e halita, o que provoca uma alta variação de propriedades desses sais na área de estudo. Observamos também que quando existe uma transição brusca da fácies carnalita para a anidrita, existe uma tendência dos métodos em classificar erroneamente a halita ao invés da anidrita, o que aparenta estar associado a fatores paleo deposicionais.

A variação no número de amostras classificadas destaca a necessidade de aprimoramentos específicos para melhorar a classificação da anidrita e, potencialmente, reduzir a diferença entre os métodos. Identificar as causas dessas variações e explorar estratégias de ajuste adequadas podem contribuir para aprimorar a precisão e a consistência dos resultados obtidos na classificação das fácies, promovendo avanços na classificação de fácies utilizando métodos de árvore de decisões. Entre os três métodos avaliados, a árvore decisão, conseguiu identificar a fácie anidrita na profundidade aproximada de 3600 metros, enquanto que o CatBoost conseguiu identificar corretamente a presença dessas fácies nas profundidades específicas de 3600 e 3850 m, associada ao maior erro de classificação. Isso destaca a capacidade de reconhecimento preciso nessas regiões mais desafiadoras, em particular para o CatBoost. Esses resultados reforçam a importância de considerar diferentes métodos e abordagens para alcançar um desempenho mais robusto e confiável na classificação das fácies.

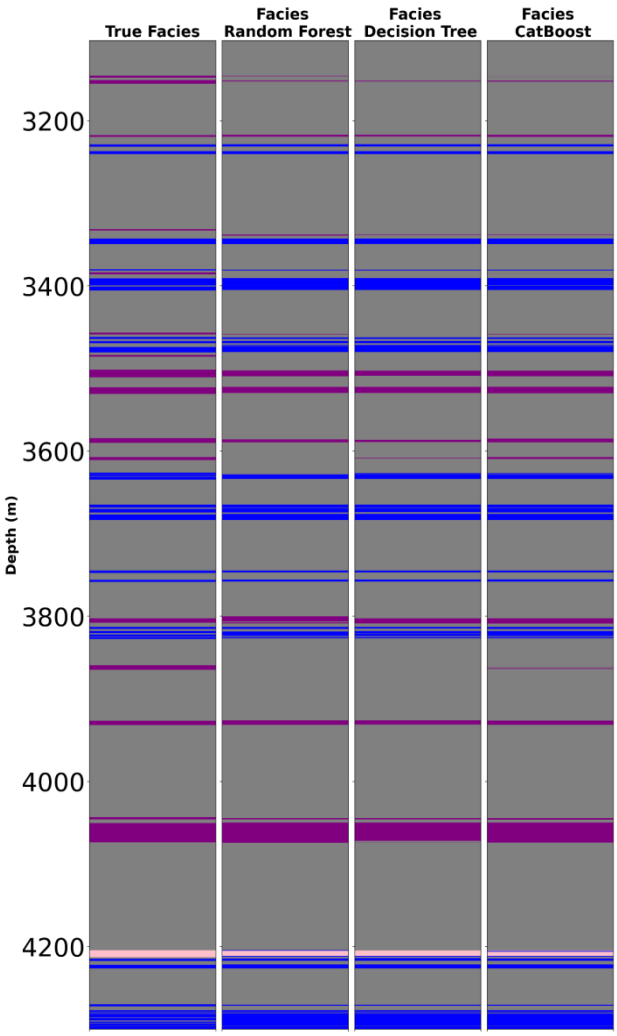

Figura 4 – *Da esquerda para a direita temos os seguintes perfis: fácies gabarito, fácies classificadas pelo Random Forest, fácies classificadas pelo Decision Tree e fácies classificadas pelo CatBoost.*

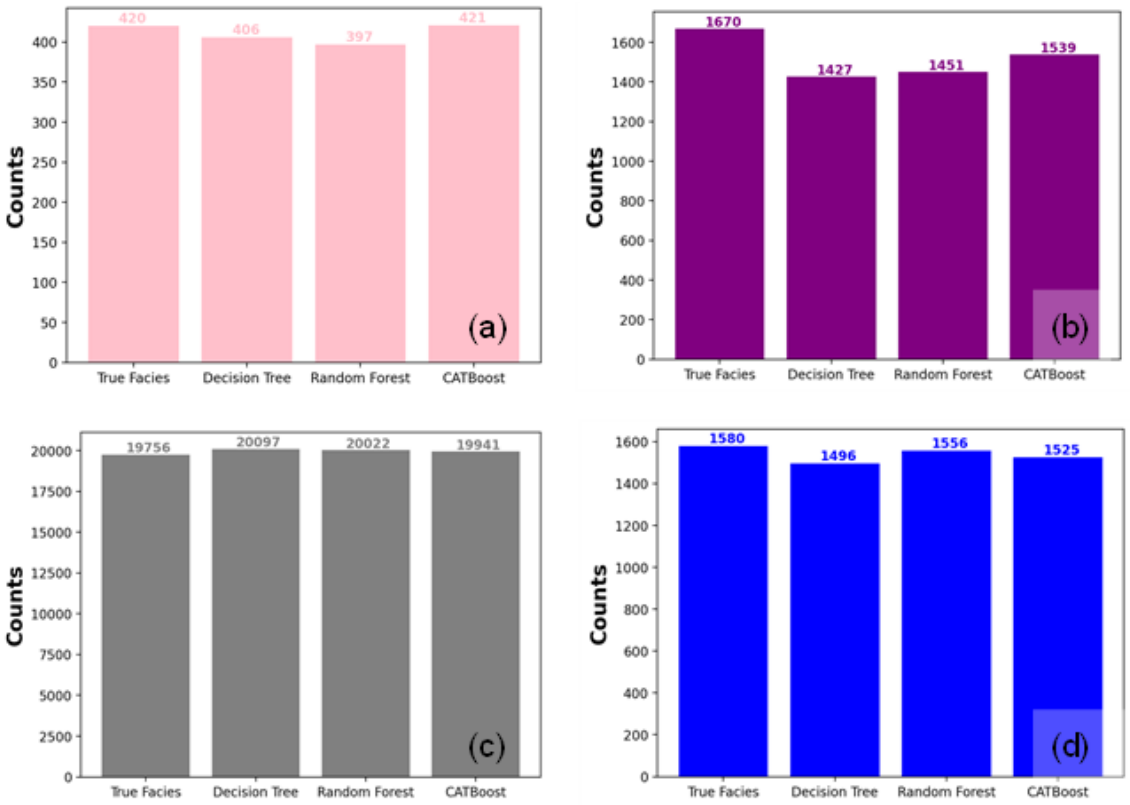

Figura 5 *- Comparação da quantidade de fácies entre o observado e o que foi classificado pelos métodos Decision Tree, Random Forest e CatBoost onde: (a) Taquidrita; (b) Anidrita, (c) Halita; (d) Carnalita.*

Conclusões

Os métodos avaliados (Random Forest, XGBoost e CatBoost) apresentaram bons resultados na classificação das fácies evaporíticas. A análise numérica mostrou que os três métodos investigados obtiveram *f1-Score* acima de 75% e acurácia acima de 85%.

Entre os métodos avaliados, o Catboost demonstrou o melhor desempenho em termos das métricas utilizadas. No entanto, quando consideramos a relação custo/benefício, observamos que o método de *Random Forest* ofereceu um resultado bem similar com um custo significativamente menor. Portanto, com base nessa perspectiva, indicamos o método de *Random Forest* para a classificação de fácies evaporíticas em grandes volumes de dados (muitos poços).

Referências

BREIMAN, LEO; FRIEDMAN, J. H.; OLSHEN, R. A.; STONE, C. J. 1984a. Classification and regression trees. Monterey, CA: Wadsworth & Brooks/Cole Advanced Books & Software. ISBN 978-0-412-04841-8

da Silva, G. F. 2022. Machine Learning aplicado na caracterização da assinatura petrofísica, espectral e geoquímica dos depósitos auríferos da Serra de Jacobina, Cráton São Francisco. Disponível em: https://repositorio.unb.br/bitstream/10482/44558/3/2022_GuilhermeFerreiradaSilva_PARCIAL.pdf

Dev, Vikrant A., Mario R. Eden. 2019. Formation Lithology Classification Using Scalable Gradient Boosted Decision Trees. Computers & Chemical Engineering. Disponível em: https://doi.org/10.1016/j.compchemeng.2019.06.001.

HALL, M.; HALL, B. 1986a. Distributed collaborative prediction: Results of the machine learning contest. The Leading Edge, Society of Exploration Geophysicists, v. 36, n. 3, p. 267–269, ISSN 1070-485X. Disponível em: https://library.seg.org/doi/10.1190/tle36030267.1

HO, T. K. 1995a. Random decision forests.. Proceedings of 3rd International Conference on Document Analysis and Recognition, Montreal, QC, Canada, pp. 278-282 vol.1, doi: 10.1109/ICDAR.1995.598994.

JOHN T. HANCOCK, TAGHI M. KHOSHGOFTAAR 2020a.“CatBoost for big data: an interdisciplinary review

LEO BREIMAN. 2001a. Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author). Statist. Sci. 16 (3) 199 - 231, August 2001. Disponível em: https://doi.org/10.1214/ss/1009213726

PAOLA GUARISO CREPALDI, ET AL. 2018a. Um estudo sobre a árvore de decisões e sua importância na habilidade de aprendizado.

QUINLAN, J.R. 1986a. Induction of decision trees. Mach Learn 1, 81–106. Disponível em: https://doi.org/10.1007/BF00116251

SALZBERG, S.L. 1993a. C4.5: Programs for machine learning by J.Ross. Morgan Kaufmann Bublishers. Disponível em: https://doi.org/10.1007/BF00993309

SAMMUT, C.; WEBB G. I. 2010a. Encyclopedia of Machine Learning.

VIZEU F 2016a. Uso de algoritmos de classificação para determinação de eletrofácies em poços da Bacia de Campos.

XICHEN, HEMANTISHWARAN 2012a. Random forests for genomic data analysis.

YAMAMOTO 2019a. Uma metodologia para a caracterização da formação Ariri utilizando dados de poços e inversão sísmica.