UNIVERSIDADE FEDERAL DE SERGIPE
PRÓ-REITORIA DE PÓS-GRADUAÇÃO E PESQUISA
COORDENAÇÃO DE PESQUISA

PROGRAMA INSTITUCIONAL DE BOLSAS DE INICIAÇÃO CIENTÍFICA – PIBIC

Caracterização de eletrofácies por meio de inteligência artificial com abordagem supervisionada

Relatório Final
Período da bolsa: de 09/23 a 08/24

Este projeto é desenvolvido com bolsa de iniciação científica

PIBIC/COPES

Sumário

1. Introdução

O processo de caracterização de eletrofácies, é a etapa de classificação dos tipos litológicos encontrados em um poço de petróleo a partir de dados de perfis elétricos. Esses perfis são propriedades físicas das rochas como radioatividade natural, densidade e resistividade, obtidas através do deslocamento contínuo de sensores de perfilagem dentro de um poço (THOMAS et al., 2001).

A execução manual desta tarefa de classificação, por parte de um geólogo, passa a ser muito exaustiva quando o número de poços a serem avaliados estão na ordem de centenas ou milhares. Logo, existe a necessidade de investir em técnicas que automatizem esse processo. Nesse contexto, algoritmos de inteligência artificial (IA) para identificação de eletrofácies, treinados com as curvas de perfis, têm sido amplamente utilizados (CARRASQUILLA, 2023; BHATTACHARYA et al., 2016).

Aprendizado de máquina é a subárea da IA que lida com as técnicas utilizadas para esse tipo de classificação. Existem duas abordagens principais para o aprendizado de máquina: supervisionada e não supervisionada. Na abordagem supervisionada, o conjunto de dados que treina o algoritmo inclui as soluções corretas, chamadas de rótulos. Já na abordagem não supervisionada, o conjunto de treinamento não possui rótulos, o que faz com que o algoritmo tente identificar padrões nos dados sem a orientação de soluções corretas (GÉRON, 2019).

Para este trabalho, escolheu-se o método de *Random Forest* (Floresta aleatória), visto que é um dos algoritmos mais utilizados na abordagem supervisionada. Esse método é baseado em árvores de decisão, estruturas capazes de dividir o conjunto de dados com base nas propriedades de uma instância, até chegar à resposta (rótulo) associada à instância em questão. O *Random Forest* gera uma coleção de árvores, da qual a previsão final é obtida através de uma votação majoritária entre as previsões das árvores que compõem o algoritmo (SHALEV-SHWARTZ; BEN-DAVID, 2014).

Um modelo de Random Forest treinado é capaz de identificar o tipo litológico associado a uma determinada instância de curvas de perfis. Cada árvore no modelo realiza divisões sucessivas dos dados, avaliando diferentes propriedades dessas curvas até chegar a uma classificação litológica correspondente. Abordagens com estudos práticos para identificação de eletrofácies, na Bacia Sergipe-Alagoas, utilizando dados de perfis de poços

e descrição de rochas, com métodos de inteligência artificial (Random Forest), não foram encontrados na literatura. Com o desinvestimento da Petrobras na Bacia Sergipe-Alagoas, os campos de petróleo desta área vêm sendo adquiridos por novas empresas produtoras que tem o interesse de torná-los viáveis do ponto de vista econômico. Nesse sentido, o desenvolvimento de ferramentas para análise das litologias presentes nesses campos pode diminuir o tempo de obtenção da informação e baixar o seu custo.

2. Objetivos

O objetivo geral deste plano de trabalho é, por meio dos perfis petrofísicos, caracterizar os tipos litológicos encontrados em poços de petróleo da Bacia Sergipe-Alagoas com o método de aprendizado de máquina conhecido como *Random Forest*.

Os objetivos específicos deste trabalho são:

I. Realizar o estado da arte sobre a identificação de eletrofácies utilizando ferramentas com métodos de inteligência artificial (Florestas Aleatórias) com abordagem supervisionada;

II. Levantar dados rotulados de perfis e definir as curvas que serão utilizadas para treinamento e teste da inteligência artificial com suporte de especialistas;

III. Escolher o algoritmo de inteligência artificial para o treinamento com os dados levantados dos poços;

IV. Realizar o teste e treinamento do algoritmo de identificação das eletrofácies a partir dos dados rotulados;

V. Obter uma inteligência artificial treinada em uma base de dados referente a algum campo de petróleo da Bacia Sergipe-Alagoas;

VI. Propagar as eletrofácies para poços que tenham as mesmas unidades estratigráficas, pertencentes, ou não, ao campo de petróleo utilizando a inteligência artificial.

Este relatório está dividido em 8 seções: Seção 1 é a introdução. Seção 2 trata dos objetivos, geral e específicos deste trabalho. A Seção 3 expõe a metodologia utilizada para obter o modelo. Seção 4 apresenta os resultados e discussões. As conclusões estão na seção

5. A seção 6 é destinada à perspectiva de trabalhos futuros. As referências bibliográficas se encontram na seção 7. Por fim, na seção 8, são apresentadas as outras atividades realizadas durante o plano de trabalho.

3. Metodologia

Esta seção apresenta a metodologia utilizada para a construção do modelo de caracterização de eletrofácies, desde a aquisição dos dados até a implementação do algoritmo. Inicialmente, são abordados o conjunto de dados, a aquisição e as principais características dos dados utilizados. Em seguida, detalha-se o processo de pré-processamento necessário para preparar os dados para o algoritmo de aprendizado de máquina. Além disso, são explicados o funcionamento e as características do método de *Random Forest.*

3.1 Conjunto de dados

No desenvolvimento deste trabalho, foram utilizados dois tipos de arquivos, o *Digital Log Interchange Standard* (DLIS) e o Arquivo Geral de Poço (AGP). Os arquivos em formato DLIS possuem os registros de perfis petrofísicos dos poços de petróleo. Os arquivos AGP são extratos digitais, em formato txt, que contém a extração dos dados litológicos utilizados. Esses dados podem ser acessados de forma gratuita por meio do Programa para Revitalização da Atividade de Exploração e Produção de Petróleo e Gás Natural em Áreas Terrestres (REATE) disponibilizado pela Agência Nacional de Petróleo, Gás Natural e Biocombustíveis (ANP) que podem ser acessado com o seguinte hiperlink: ANP-TERRESTRE (cprm.gov.br). A tabela 1 mostra os poços utilizados no estudo com as suas nomenclaturas ANP e da PETROBRAS que os nomeou por campo.

Tabela 1 - Nomes dos poços

Nome ANP	Nome BR	Nome abreviado
1-BRSA-459-SE	1-NCL-2-SE	P_459
1-BRSA-551-SE	1-FSG-1-SE	P_551
1-BRSA-574-SE	1-FSJQ-1-SE	P_574
1-BRSA-595-SE	1-FSG-2-SE	P_595
1-BRSA-605-SE	1-FSJQ-2-SE	P_605
1-BRSA-643-SE	1-CBO-1-SE	P_643
1-BRSA-645-SE	1-SIB-1-SE	P_645
1-BRSA-659-SE	1-POI-2-SE	P_659
1-BRSA-689-SE	1-POI-1-SE	P_689
1-BRSA-696-SE	1-FSJQ-3-SE	P_696
1-BRSA-698-SE	1-DP-2-SE	P_698

Fonte: Autor

3.1.1 Curvas de perfis petrofísicos

Para o trabalho em questão, foram escolhidas 9 curvas de perfis petrofísicos, assim como no trabalho de SOUSA (2024):

I. Raios Gama (GR): Detecta a radioatividade natural total da formação geológica. É mais utilizada para identificação da litologia e cálculo do teor de argila nas litologias. As unidades são em °API.

II. Porosidade Neutrônica (NPHI): As versões mais avançadas desta ferramenta medem a quantidade de nêutrons epitermal e/ou termal da rocha após o bombardeio de nêutrons de alta energia emitidos de uma fonte radioativa na ferramenta. Sua principal utilização é na estimativa da porosidade da rocha a partir do conteúdo de hidrogênio no meio e identificação da presença de gás.

III. Resistividade Profunda (RESD): Oferece uma leitura aproximada da resistividade, através da medição de campos elétricos e magnéticos induzidos nas rochas ou através de correntes elétricas focalizadas para dentro das camadas. É utilizada principalmente na identificação dos fluidos (água ou hidrocarbonetos)

presentes nas litologias e, secundariamente, no tipo de litologia. Unidades: ohm·m.

IV. Sônico (DT): Mede os tempos de trânsito entre dois detectores, localizados a uma distância fixa na ferramenta, de uma onda mecânica compressional. É utilizado essencialmente na obtenção da porosidade das rochas e o tempo de trânsito dessa onda. Permite a conversão do tempo em profundidade nas seções sísmicas. As unidades são em microssegundo por pé.

V. Densidade (RHOB): O Perfil de Densidade registra continuamente as variações das densidades das camadas (g/cm^3). Essa densidade é mensurada através do bombardeio das camadas por um feixe monoenergético de raios gama emitidos por uma fonte radioativa (Cs137), direcionada e pressionada contra a parede do poço para minimizar os efeitos do poço (lama e reboco). É utilizada para a obtenção da porosidade da rocha. Unidades: gramas por centímetro cúbico.

VI. Compensação do Perfil de Densidade (DRHO): Esta curva é adquirida em conjunto com a curva de densidade. Serve como controle de qualidade da mesma. Mostra a correção que foi adicionada à curva de RHOB ao longo da perfilagem devido a efeitos do poço (reboco e arrombamento). Unidades: gramas por centímetro cúbico.

VII. Fator Fotoelétrico (PE): Esta curva é adquirida em conjunto com a curva de densidade. O efeito fotoelétrico é proveniente da absorção total dos raios gama de baixa energia pelos elétrons e possui relação direta com o número atômico do elemento químico que está compondo a rocha. O raio gama incidente apresenta baixo nível de energia após o choque com os átomos da formação, sendo totalmente absorvido pelo átomo durante a colisão e transmitindo sua energia cinética para o elétron, ejetando-o. É utilizado frequentemente na identificação litológica. Unidades: B/e (Barns/électron - unidade de absorção de fator fotoelétrico).

VIII. Caliper (CAL): Esta curva é adquirida, geralmente, em conjunto com a curva de densidade e mede o diâmetro do poço e a "rugosidade" da parede do mesmo. É utilizada como controle de qualidade dos perfis corridos permitindo identicar os trechos com grandes desmoronamentos e que pode ter afetado as medidas das demais curvas. Unidades: polegadas.

IX. Diâmetro da Broca (BS): Essa curva nos fornece o diâmetro da broca que foi usada para perfurar o poço. Unidades: polegadas.

3.1.2 Unidades Geológicas

Além das curvas de perfis, as unidades geológicas onde se encontravam os tipos litológicos, também desempenharam o papel de variável preditiva. Duas colunas estratigráficas, Unidade Formação e Unidade Membro, foram retiradas dos arquivos AGP de cada poço. A fim de obter uma representação única para cada unidade na hierarquia mais baixa possível para cada unidade geológica, realizou-se uma padronização de nomenclatura para essa variável, conforme apresentado na tabela 2.

Tabela 2 - Padronização do nome das unidades.

Nome da Unidade Formação	**Formação extraída do AGP**	**Nome da Unidade Membro**	**Membro extraído do AGP**	**Nome Padrão das Unidades**
Barreiras	BARRRS	_	_	BAR_BAR
Cotinguiba	COTING	Sapucari	SAPUCA	COT_SAP
Riachuelo	RIACHU	Angico	ANGICO	RIA_ANG
Riachuelo	RIACHU	Maruim	MARUIM	RIA_MAR
Riachuelo	RIACHU	Taquari	TAQUAR	RIA_TQR
Muribeca	MURIBE	Oiteirinhos	OITEIR	MUR_OIT
Muribeca	MURIBE	Ibura	IBURA	MUR_IBU
Muribeca	MURIBE	Carmópolis	CARMOP	MUR_CPS
Penedo	PENEDO	_	_	PDO_PDO
Serraria	SERRAR	_	_	SER_SER
Bananeiras	BANANE	_	_	BAN_BAN
Aracaré	ARACAR	_	_	ARA_ARA
Batinga	BATING	_	_	BAT_BAT
Embasamento	EMBASA	_	_	EMB_EMB

Fonte: Autor

3.1.3 Tipos Litológicos

O foco central deste trabalho é a caracterização, por meio dos perfis petrofísicos, das eletrofácies, que é o termo utilizado para se referir aos tipos litológicos interpretados a partir das propriedades físicas obtidas nos perfis. Dentre os 11 poços que foram utilizados para treinar o modelo de inteligência artificial, existiam 11 tipos litológicos em seus respectivos arquivos AGP: Anidrita, Arenito, Calcarenito, Calcilutito, Conglomerado, Dolomito, Filito, Folhelho, Granito, Siltito e Xisto. Dessa maneira, a fim de atingir o objetivo geral

estabelecido, o modelo foi treinado de modo a ser capaz de receber como entrada valores de curvas de perfis e categorizar o tipo litológico respectivo, como uma dessas 11 classes.

3.2 Tratamento dos dados

Antes de aplicar o algoritmo de Floresta Aleatória, é fundamental que os dados passem por um processo de tratamento para que eles possam ser analisados pelos especialistas e utilizados pelo algoritmo. Dessa forma, o tratamento foi realizado por meio do seguinte procedimento:

I. Conversão da unidade de medida de profundidade de polegada para metro; (seção 3.2.1)

II. Remoção de registros que não possuíam todas as curvas escolhidas; (seção 3.2.2)

III. Padronização do nome das curvas; (seção 3.2.3)

IV. Adição da curva DCAL em poços que ainda não a possuem; (seção 3.2.4)

V. Reclassificação das litologias que estão fora dos padrões estabelecidos. (seção 3.2.5)

3.2.1 Conversão da unidade de medida de profundidade

Para manter o padrão usual da indústria de petróleo do Brasil, a unidade de medida de profundidade de 7 dos 11 poços passou por um processo de conversão de polegada (pol) para metro (m). Os poços P_459, P_551, P_605, P_643, P_689, P_696 e P_698 passaram por essa etapa do processamento, visto que possuíam a unidade de medida de profundidade em polegada.

3.2.2 Remoção dos registros que não possuem todas as curvas escolhidas

Em determinados poços, o arquivo dlis organizava os registros de perfilagem em diversos quadros de dados (*frames*), com diferentes intervalos de profundidade. Dentre esses diversos intervalos, alguns não possuíam registro de todas as curvas de perfis petrofísicos escolhidas para o trabalho. Para garantir que o algoritmo seja treinado com todo o conjunto de variáveis definido, foi realizada uma filtragem, preservando apenas as amostras que possuíam todas as curvas requisitadas para o estudo. A tabela 3 mostra quais poços possuíam mais de um *frame* e a quantia de *frames* utilizados, que continham todas curvas escolhidas.

Tabela 3 – Número de frames em cada poço

Poço	**N° de *frames***	**N° de *frames* utilizados**
P_459	9	4
P_551	9	4
P_574	1	1
P_595	1	1
P_605	4	4
P_643	5	3
P_645	1	1
P_659	1	1
P_689	9	4
P_696	9	4
P_698	9	4

Fonte: Autor

3.2.3 Padronização do nome das curvas

De acordo com o especialista Luiz Henrique Vandelli, existem inúmeros mnemônicos de curvas de resistividade profunda que poderiam ser utilizados no estudo. Conforme acontece em companhias de perfilagem, existem dois princípios de funcionamento das ferramentas de resistividade: indução e laterolog. O princípio da indução engloba as curvas ILD, RILD, IEL, AIT90, AHT90, RT90, AT90, AO90, RT, AF90, AHF90 e AFH90. Já o laterolog possui as curvas LLD, RLLD, HDRS, HLLD, LL7 e RLL7.

Assim como acontece com os perfis de resistividade profunda, existem curvas equivalentes para outros perfis utilizados, com uma variação de nomenclatura. Desse modo, a fim de padronizar os nomes dessas variáveis, duas ou mais curvas equivalentes tiveram seus nomes renomeados para um nome padrão escolhido. No caso da resistividade, foi acordado utilizar o termo RESD para se referir a uma curva de resistividade profunda.

Na tabela 4 estão apresentados os conjuntos de curvas equivalentes e o nome padrão que foi definido para ser utilizado neste trabalho.

Tabela 4 – Nomes equivalentes e nome padrão escolhido para cada perfil

Nome do perfil	Nomes equivalentes	Nome padrão escolhido
Raios Gama	GR	GR
Porosidade Neutrônica	NPHI	NPHI
Resistividade Profunda	ILD; RILD; IEL; AIT90; AHT90; RT90; AT90; AO90; RT; AF90; AHF90; AFH90; LLD; RLLD; HDRS; HLLD; LL7; RLL7	RESD
Sônico	DT; DTC	DT
Densidade	RHOB; RHOZ	RHOB
Compensação da Densidade	DRHO; HDRA	DRHO
Fator Fotoelétrico	PE; PEFZ; PEU	PE
Caliper	HCAL; CAL; CALI	CAL
Diâmetro da Broca	BS; BSZ	BS

Fonte: Autor

Em alguns poços foram identificadas múltiplas curvas de resistividade profunda. Nesses casos, foi imprescindível manter apenas uma curva e eliminar as demais, a fim de evitar que a duplicação dessa característica influenciasse negativamente o processo de aprendizado do modelo. Nos casos em que a filtragem foi necessária, priorizou-se a manutenção da curva RT, com a subsequente remoção das demais.

3.2.4 Adição da curva DCAL

Para fins de controlar a qualidade dos dados utilizados, foi adicionado o perfil de controle DCAL, que é o resultado da diferença entre o valor de CAL e BS. Por meio do DCAL é possível identificar se uma determinada região de um poço tem reboco (DCAL negativo) ou arrombamento (DCAL positivo). Para evitar que regiões que possuam elevado grau de reboco ou arrombamento impactem negativamente o modelo, os registros de dados utilizados neste trabalho possuem o DCAL dentro do intervalo [-1, 1,5], em polegadas.

3.2.5 Reclassificação das litologias fora dos padrões estabelecidos

Segundo o especialista Vitor Hugo Simon, geólogo consultor do projeto, as litologias presentes no Embasamento Fraturado da Sub-bacia de Sergipe, na região do Alto de Aracaju

são, principalmente, rochas metamórficas (xistos e filitos) e secundariamente rochas graníticas.

Durante o pré-processamento dos dados, foi identificado no poço P_459, um tipo de litologia descrita como “metamórfica não identificada”, referente ao Embasamento da Sub-bacia de Sergipe. Sendo esse um termo muito genérico para classificar rochas e com a intenção de obter uma caracterização mais precisa das litologias, as amostras de “metamórfica não identificada” foram submetidas a um algoritmo de classificação litológica para se obter uma caracterização mais precisa do tipo litológico.

Para realizar a reclassificação, foi utilizado um modelo de classificação de floresta aleatória, treinado com as propriedades de perfis das amostras litológicas de xisto e filito, presentes nos poços P_689 e P_659. Após o treinamento, o modelo foi capaz de rotular as amostras de “metamórfica não identificada” como uma dessas duas categorias.

No conjunto de dados em estudo, havia 540 amostras de xisto e 202 de filito, nos poços P_689 e P_659, respectivamente. Dessa maneira, o algoritmo de classificação foi treinado e testado com um total de 742 registros de xisto e de filito.

A divisão das amostras em treinamento e teste foi realizada separando 2/3 do total para treinamento e 1/3 para teste, conforme recomendado na literatura (FACELI et al., 2011). Os resultados da divisão são apresentados na tabela 5.

Tabela 5 - Divisão das amostras de xisto e filito em conjuntos de treino e teste

Tipo Litológico	**Número de amostras**	**Amostras para treinamento**	**Amostras para teste**
Filito	202	132	70
Xisto	540	362	178
Total	742	494	248

Fonte: Autor

Uma vez que o algoritmo foi treinado, a sua acurácia pôde ser testada com o conjunto de teste. Isto é, após aprender com as amostras para treinamento, o modelo foi testado ao classificar, em xisto ou filito, a litologia correspondente aos 248 registros de propriedades de perfis petrofísicos separados para teste. A acurácia do modelo foi de 99.59%.

Em seguida, após as etapas de treinamento e teste, o algoritmo foi utilizado para

categorizar as amostras de “metamórfica não identificada” em xisto ou filito, com base em suas respectivas propriedades. Dentre os 585 registros de metamórfica que existiam no conjunto de dados, 573 foram reclassificados como xisto e apenas 12 como filito.

De acordo com os especialistas, não foi observado, até então, na unidade EMB, a existência de mais de um tipo litológico. Desse modo, todos os registros de “metamórfica não identificada” foram convertidos em xisto, e os filitos foram considerados como erros de classificação.

3.3 Algoritmo

Neste trabalho foi empregado o método de *random forest*, algoritmo baseado em árvore com abordagem supervisionada, para caracterizar eletrofácies. Um modelo *random forest* de classificação é uma coleção de árvores de decisão, da qual a previsão final é obtida através de uma votação majoritária entre as previsões realizadas por cada uma das árvores individuais que compõem o algoritmo (SHALEV-SHWARTZ; BEN-DAVID, 2014). Desse modo, a classe mais frequentemente prevista pelas árvores será a previsão final do modelo.

3.3.1 Variáveis preditivas

Variáveis preditivas, também chamadas de *features* ou atributos, são os dados de entrada usados em um modelo de aprendizado de máquina. Essas variáveis contêm informações que o algoritmo analisa para identificar padrões e gerar uma saída, isto é, a previsão.

Os perfis petrofísicos escolhidos para desempenharem papel de variável preditiva no treinamento do modelo, são aqueles que, segundo os especialistas, são discriminatórios:

I. GR;
II. RESD;
III. DT;
IV. RHOB;
V. DRHO;
VI. NPHI;
VII. PE.

Além das curvas de perfis, também foram incluídas as unidades geológicas no

conjunto de variáveis preditivas, para alguns dos modelos avaliados. Um modelo de floresta aleatória não possui a capacidade de processar diretamente variáveis categóricas, ou seja, dados que não são numéricos, como o nome de uma unidade. Uma solução comum para corrigir esse problema é criar um atributo binário para cada categoria. Esse método é denominado codificação One-Hot Encoding, pois apenas um dos atributos terá o valor 1 (ativo), enquanto os demais permanecerão com o valor 0 (inativos) (GÉRON, 2019). Desse modo, o *One-Hot Encoding* transforma uma variável categórica em várias colunas binárias, onde os valores assumem zero ou um, sendo cada nova coluna respectiva a uma Unidade.

Para cada observação de uma Unidade no conjunto de dados, sua coluna correspondente é marcada como 1, enquanto as demais colunas são marcadas como 0. Isso permite que o modelo trate cada categoria como uma entidade independente, sem introduzir uma relação de ordem que poderia causar problemas em modelos de *Random Forest*. A tabela 6 ilustra esse procedimento por meio de um exemplo prático, no qual a primeira linha possui a Unidade MUR_CPS, a segunda linha corresponde a RIA_ANG e a terceira a BAN_BAN.

Tabela 6 - Exemplo de colunas binárias geradas pelo processo de *One-Hot Encoding*

ARA ARA	BAN BAN	BAR BAR	BAT BAT	COT SAP	MUR CPS	MUR IBU	MUR OIT	PDO PDO	RIA ANG	RIA MAR	RIA TQR	SER SER
0	0	0	0	0	**1**	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	**1**	0	0	0
0	**1**	0	0	0	0	0	0	0	0	0	0	0

Fonte: Autor

3.3.2 Divisão do conjunto de dados

A fim de aplicar o algoritmo de Random Forest, o conjunto total de dados foi divido em duas partes principais: conjunto de treinamento e conjunto de teste. Escolheu-se a proporção de divisão usualmente utilizada de 2/3 das amostras para treinamento do algoritmo e 1/3 para teste (FACELI et al., 2011). O conjunto de treinamento é fornecido ao modelo de *Random Forest* para que ele possa aprender padrões e relacionamentos dentro dos dados. Na etapa de teste, o modelo treinado recebe os dados do conjunto de teste para fazer previsões em novos dados não vistos (TCHAKOUCHT et al., 2024). A tabela 7 apresenta os números da divisão do conjunto de dados deste estudo por tipo litológico.

Tabela 7 - Número de amostras para treinamento e teste do modelo

Tipo Litológico	Número de amostras	Amostras para Treinamento	Amostras para Teste
Anidrita	1262	849	413
Arenito	8508	5720	2788
Calcarenito	1828	1246	582
Calcilutito	6302	4117	2185
Conglomerado	892	575	317
Dolomito	361	245	116
Filito	202	137	65
Folhelho	15859	10595	5264
Granito	445	289	156
Siltito	516	337	179
Xisto	1125	756	369
Total	37300	24866	12434

Fonte: Autor

4. Resultados e discussões

Esta seção é destinada à análise dos resultados obtidos com o modelo de aprendizado de máquina para caracterização de eletrofácies. Neste trabalho foram treinados três modelos diferentes de classificação, com as seguintes características de dados de entrada:

I. Dados desbalanceados e sem unidades geológicas como variáveis preditivas; (seção 4.1)

II. Dados desbalanceados e com unidades geológicas como variáveis preditivas; (seção 4.2)

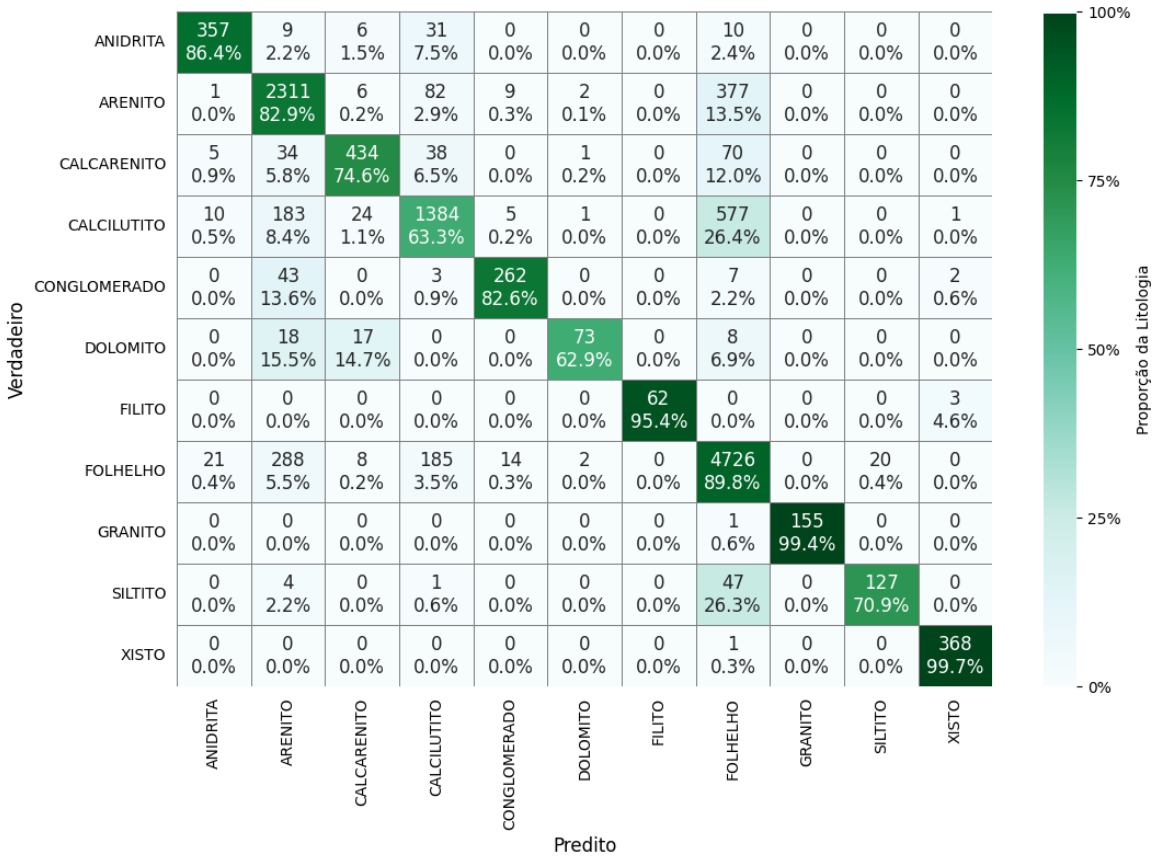
III. Dados balanceados e com unidades geológicas como variáveis preditivas. (seção 4.3)

4.1 Modelo 1 - Dados desbalanceados e sem unidades geológicas

O primeiro modelo foi treinado apenas com as curvas de perfis petrofísicos discriminatórios como variáveis preditivas, ou seja, não se utilizou das unidades geológicas na etapa de treinamento. Ademais, os dados não foram balanceados, de modo que determinados tipos litológicos tiveram um número elevado de amostras de treino e teste.

A figura 1 é a matriz de confusão do primeiro modelo treinado, do qual teve seu desempenho avaliado com o seu conjunto de teste. No eixo y encontram-se as classes litológicas verdadeiras e no eixo x as preditas. Cada célula da matriz possui dois valores: a quantidade absoluta e a porcentagem de amostras da litologia da linha (verdadeira) que foram classificadas como a litologia da coluna (predita).

Figura 1 - Matriz de confusão do teste do primeiro modelo



Fonte: Autor

O modelo obteve uma acurácia geral de 82,50%, como pode ser observado na figura 1, no qual o número de acertos se concentrou na diagonal principal da matriz. Nove das onze classes tiveram mais de 70% de acerto, apenas Calcilutito e Dolomito obtiveram 63,3% e 62,9% de acerto, respectivamente. O desempenho das previsões de Calcarenito e Siltito, se manteve entre 70% e 80%, enquanto as classes Anidrita, Arenito, Conglomerado e Folhelho preencheram o intervalo de 80% a 90%. Os três tipos litológicos do Embasamento, Filito, Granito e Xisto, obtiveram uma porcentagem de acerto acima de 90%.

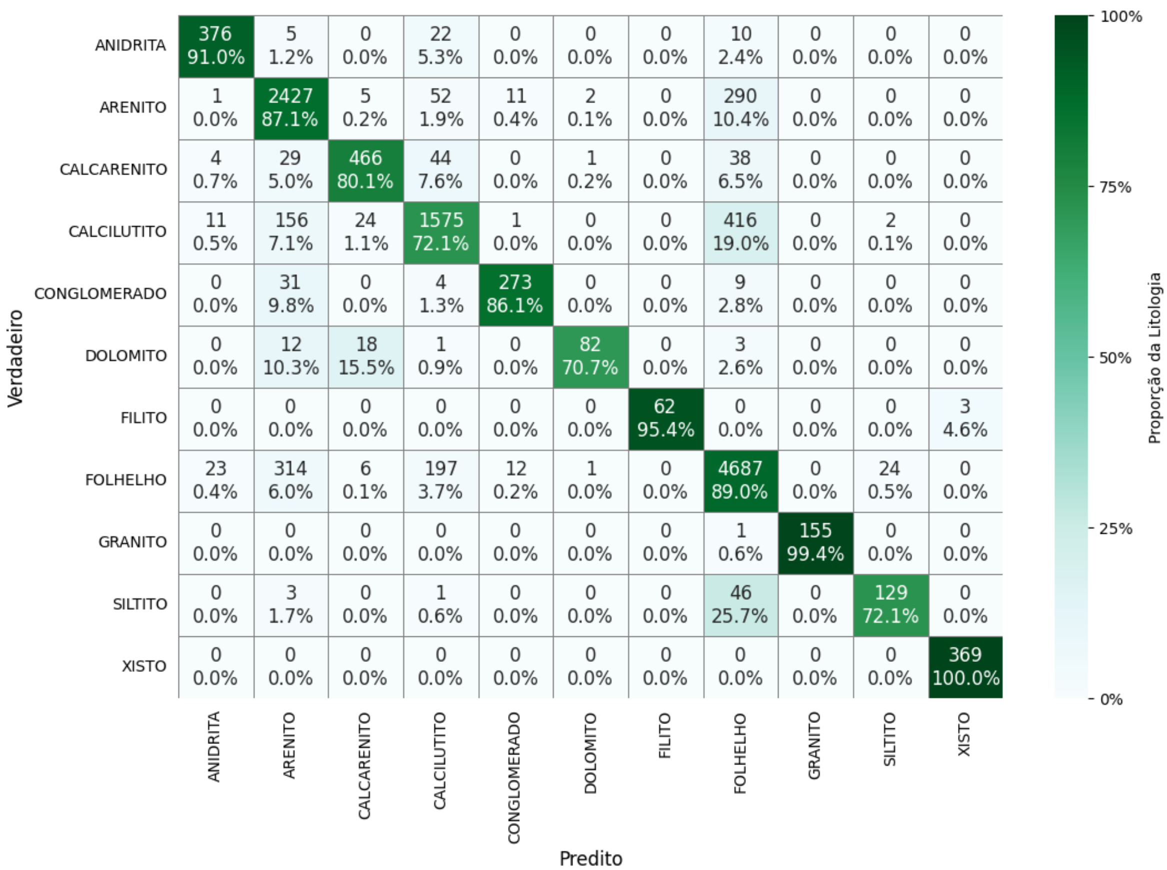
As amostras de teste se confundiram, principalmente, com a classe litológica mais

frequente, o Folhelho. As classes Calcilutito e Siltito são as que mais se confundem com 26,4% e 26,3%, respectivamente, das amostras caracterizadas como Folhelho. De maneira semelhante, o Arenito, segunda classe mais numerosa, gerou mais de 10% de confusão no Conglomerado e Dolomito, tipos litológicos menos frequentes. Diante disso, no modelo seguinte foram introduzidas as unidades geológicas.

4.2 Modelo 2 - Dados desbalanceados e com unidades geológicas

Em segunda análise, o próximo modelo avaliado, se utiliza, além dos perfis geofísicos, das unidades geológicas para fazer previsões acerca do tipo litológico. Visto isso, além dos sete perfis que treinaram o primeiro modelo, as onze variáveis binárias referentes às unidades com nomes padronizados, apresentadas na seção 3.3.1 participaram do aprendizado. A figura 2 é a matriz de confusão referente ao teste deste modelo.

Figura 2 - Matriz de confusão do teste do segundo modelo



Fonte: Autor

A adição das unidades geológicas otimizou o desempenho do algoritmo, de modo que a acurácia geral deste modelo passou para 85,25%, 2,75% a mais quando comparado com o anterior. Por meio dessa melhora, pode ser observado uma porcentagem de acerto acima de 70% em toda a diagonal principal da matriz de confusão.

A tabela 8 apresenta a comparação do número de acertos por litologia entre os dois modelos e a variação do primeiro para o segundo. Conforme exposto na coluna da variação, o aprendizado com as unidades geológicas trouxe uma melhora no desempenho do algoritmo em sete das onze classes litológicas. Apenas Filito e Granito, com zero, e Folhelho, com uma variação negativa de 0,8%, não apresentaram melhora. A precisão quanto as demais classes apresentou um aumento que variou entre 0,3% (Xisto) e 8,8% (Calcilutito).

Tabela 8 - Comparativo de acertos por litologia entre os dois modelos

Tipo Litológico	Acertos Modelo 1 (%)	Acertos Modelo 2 (%)	Variação (%)
Anidrita	86,4	91,0	+4,6
Arenito	82,9	87,1	+4,2
Calcarenito	74,6	80,1	+5,5
Calcilutito	63,3	72,1	+8,8
Conglomerado	82,6	86,1	+3,5
Dolomito	62,9	70,7	+7,8
Filito	95,4	95,4	0
Folhelho	89,8	89,0	-0,8
Granito	99,4	99,4	0
Siltito	70,9	72,1	+1,2
Xisto	99,7	100,0	+0,3

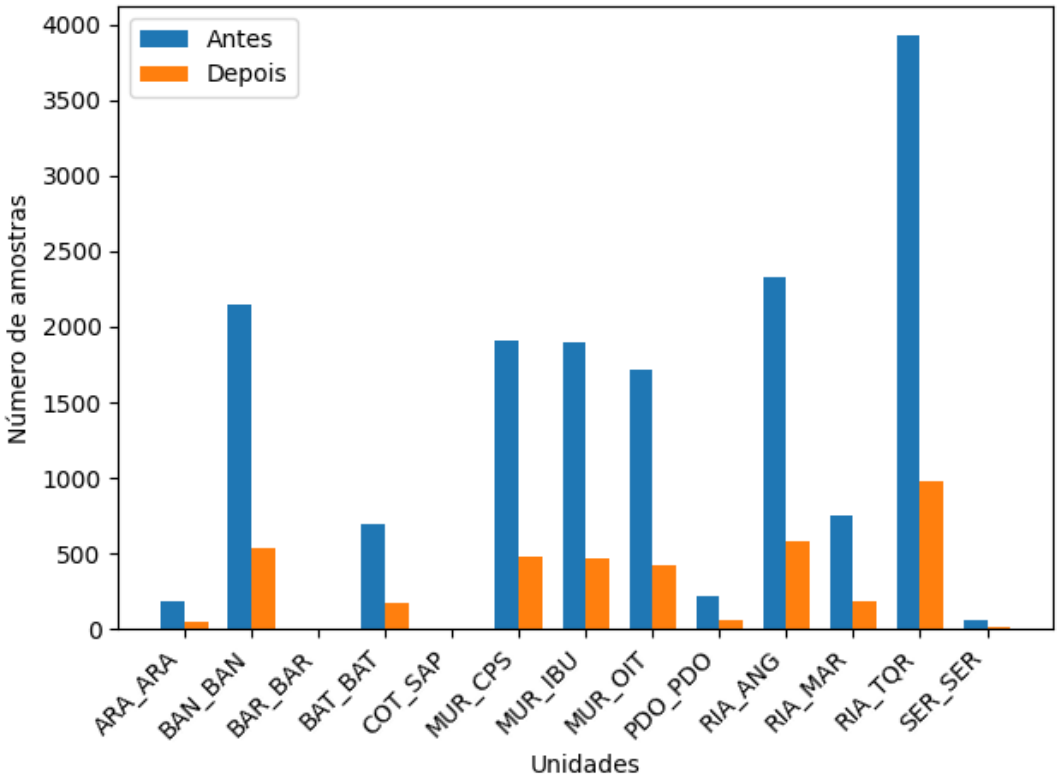
Fonte: Autor

4.3 Modelo 3 - Dados balanceados e com unidades geológicas

O tipo litológico mais frequente no conjunto de dados, Folhelho, induziu 1098 e 813 predições erradas, no primeiro e segundo modelo, respectivamente. Visto isso, foi realizado um balanceamento nos dados, com um método comum para lidar com problemas de desequilíbrio entre classes: o *undersampling*, do qual consiste em remover instâncias da classe majoritária (PEDRAJAS, 2024). De modo a balancear os dados de maneira proporcional em todas as unidades geológicas, foi removido de maneira aleatória em cada

unidade, 2/5 dos registros de Folhelho. A figura 3 apresenta o número de amostras de Folhelho por unidade geológica, antes e depois do balanceamento.

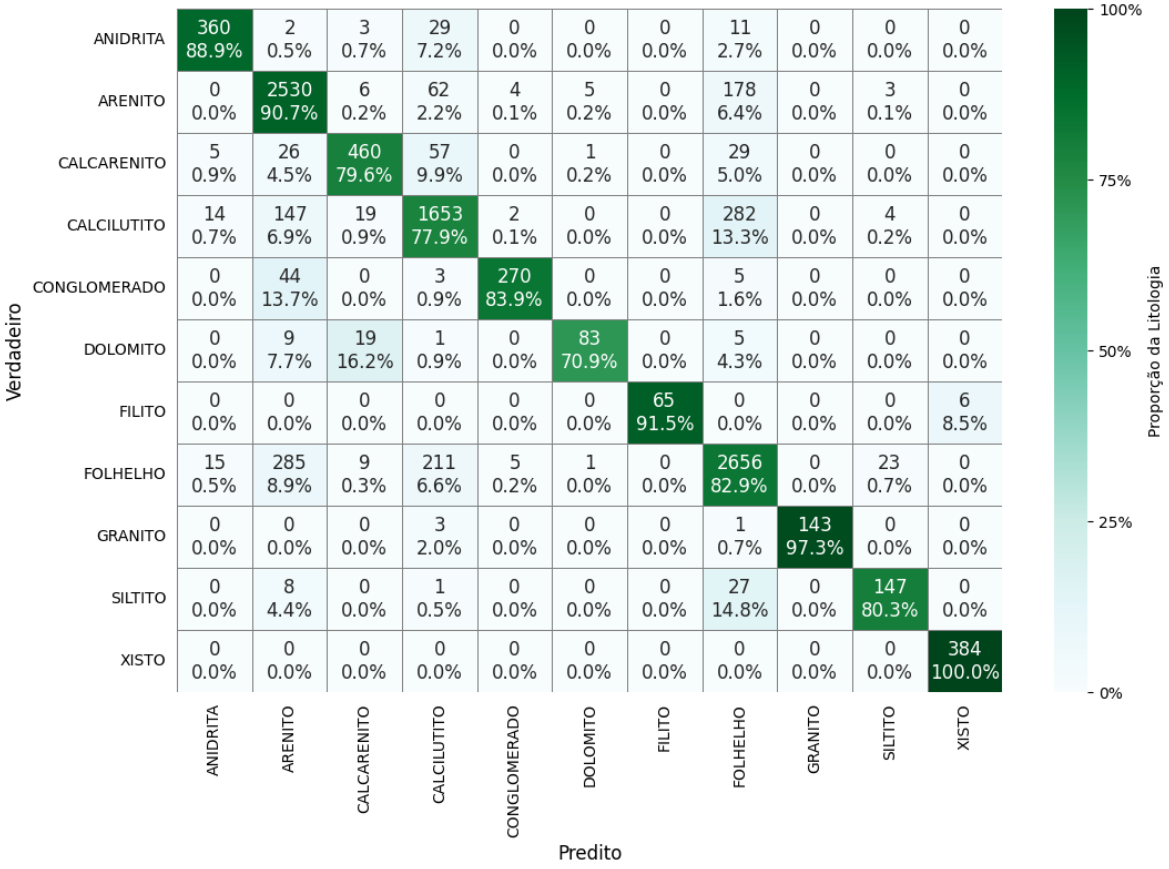
Figura 3 - Número de amostras de Folhelho por unidade geológica, antes e depois do balanceamento



Fonte: Autor

O treinamento deste terceiro modelo utilizou as mesmas variáveis preditivas do segundo, isto é, os sete perfis petrofísicos definidos em e as unidades geológicas apresentadas na seção 3.3.1. A matriz de confusão deste modelo é mostrada na figura 4.

Figura 4 - Matriz de confusão do teste do terceiro modelo



Fonte: Autor

A acurácia geral deste modelo foi similar ao anterior, 85.56%. A diferença chave entre os dois modelos que englobam as unidades geológicas, foi a quantidade de amostras confundidas com Folhelho, que diminuiu após o balanceamento. A tabela 9, mostra a comparação do número de amostras de cada tipo litológico confundidas com Folhelho, antes e depois de balancear os dados.

Tabela 9 - Comparativo de amostras confundidas com Folhelho

Tipo Litológico	Confusão Modelo 2 (%)	Confusão Modelo 3 (%)	Variação (%)
Anidrita	2,4	2,7	+0,3
Arenito	10,4	6,4	-4,0
Calcarenito	6,5	5,0	-1,5
Calcilutito	19,0	13,3	-5,7
Conglomerado	2,8	1,6	-1,2
Dolomito	2,6	4,3	+1,7
Filito	0,0	0,0	0,0
Granito	0,6	0,7	+0,1
Siltito	25,7	14,8	-10,9
Xisto	0,0	0,0	0,0

Fonte: Autor

De acordo com o exposto na coluna da variação, a confusão com Folhelho foi reduzida em cinco das dez classes litológicas. O restante se manteve constante ou aumentou entre 0.1% e 1.7%. As maiores reduções se encontram nas classes Arenito, Calcilutito e Siltito: -4.0%, -5.7% e -10.9% respectivamente. Os resultados obtidos mostram a superioridade deste último modelo em relação aos demais e a importância do balanceamento dos dados.

4.4 Comparação entre os modelos

Conforme os 3 modelos apresentados, observa-se uma melhora na precisão da classificação das eletrofácies com a incorporação de técnicas como a adição das unidades geológicas e o balanceamento dos dados. A Tabela 10 apresenta a média e o desvio padrão dos 3 modelos avaliados. A média de desempenho do algoritmo nos 11 tipos litológicos aumentou progressivamente do modelo 1 para o modelo 3. Em contrapartida, o desvio padrão foi reduzido ao longo da mesma transição, o que se deve ao fato de o algoritmo ter se tornado mais uniforme na classificação de todas as litologias, resultado da inclusão das unidades geológicas e do balanceamento dos dados.

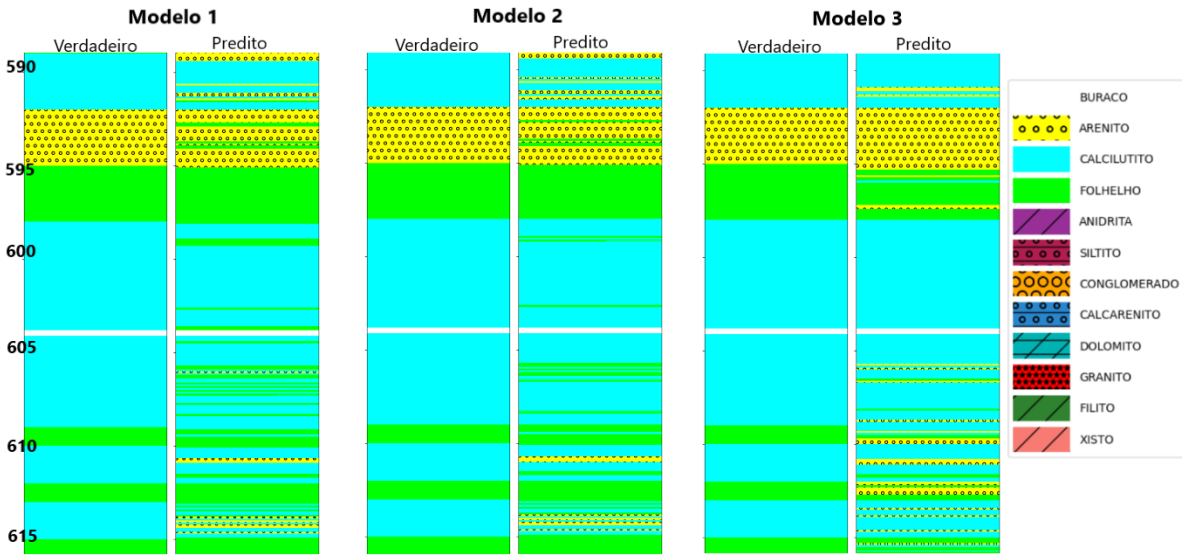
Tabela 10 – Comparação entre os modelos avaliados

Medida	Modelo 1	Modelo 2	Modelo 3
Média	82,54	85,72	85,80
Desvio Padrão	12,67	10,24	8,34

Fonte: Autor

A figura 5 é uma visualização gráfica comparativa entre os tipos litológicos verdadeiros e preditos ao longo do poço P_551 nos três modelos avaliados. Cada modelo tem a coluna de classes litológicas verdadeiras à esquerda e as preditas à direita. Este recorte do poço P_551 inclui o intervalo de profundidade de 590m a 615m.

Figura 5 – Comparação gráfica dos modelos avaliados (recorte do poço P_551)



Fonte: Autor

De acordo com o conteúdo apresentado na Figura 5, assim como nas matrizes de confusão dos modelos, observa-se uma diminuição no número de amostras de Calcilutito confundidas com Folhelho, após a inclusão das unidades geológicas. Essa redução tornou-se ainda mais significativa a partir do modelo 3, como consequência do balanceamento da classe Folhelho.

5. Conclusões

Neste trabalho, foram propostos modelos de inteligência artificial para o problema de caracterização de eletrofácies, tradicionalmente abordado de forma manual por especialistas da área da Geologia. Por meio da aquisição e pré-processamento dos dados, foi possível fornecer ao algoritmo de *random forest*, recursos o suficiente para capacitá-lo a realizar previsões acerca da identificação dos tipos litológicos. Os modelos treinados apresentaram, na maioria dos casos observados, taxas de acerto por litologia variando entre 70% e 100%.

Foi possível notar, através da comparação entre modelos, que a prática de incluir as unidades em nível hierárquico mais baixo no processo de treinamento do algoritmo, contribuiu para a precisão da classificação em estudo. Assim sendo, conclui-se que as unidades geológicas possuem informações valiosas, que podem ser utilizadas pelo processo de aprendizado de máquina, para melhor discriminar os tipos litológicos em estudo.

Ademais, observou-se que, embora haja uma perda de informações decorrente do processo de *undersampling*, o balanceamento dos dados, especialmente em classes majoritárias, pode reduzir a confusão gerada nas demais classes menos numerosas. Portanto, é válido avaliar a possibilidade de reduzir o número de amostras, dos tipos litológicos mais frequentes, no conjunto de poços que se deseja trabalhar, a fim de que o modelo aprenda e favoreça todas as classes de maneira mais igualitária.

6. Perspectivas de futuros trabalhos

Por meio deste documento, foi exposto a utilização de 7 curvas de perfis para classificar tipos litológicos. Para o próximo ciclo do programa PIBIC 24/25, tem-se como objetivo, a introdução de perfis especiais de imagens para a identificação de eletrofácies. Os perfis de imagens acústicas e resistivas apresentam uma grande quantidade de dados por profundidade, o que pode enriquecer os modelos para uma melhor predição das litologias.

Em relação ao trabalho focado nas ferramentas convencionais, visa-se desenvolver técnicas para analisar poços profundos, dos quais possuem mais de uma corrida de perfil. Outra possibilidade de continuidade é definir quais as curvas são mais relevantes para serem utilizadas no treinamento dos algoritmos. Ademais, espera-se trabalhar com algoritmos que não necessitam do conjunto completo de curvas, de modo que seja possível dar utilidade a amostras de profundidade que não possuem todas as curvas registradas.

7. Referências bibliográficas

BHATTACHARYA, S. et al. Comparison of supervised and unsupervised approaches for mudstone lithofacies classification: Case studies from the Bakken and Mahantango-Marcellus Shale, USA, 2016.

CARRASQUILLA, A. Lithofacies prediction from conventional well logs using geological information, wavelet transform, and decision tree approach in a carbonate reservoir in southeastern Brazil, 2023.

FACELI, K. et al. Inteligência Artificial: Uma Abordagem de Aprendizado de Máquina. [S.l.]: LTC, 2011.

GÉRON, A. Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems. [S.l.]: O'Reilly Media, 2019.

PEDRAJAS, N. Partial random under/oversampling for multilabel problems, 2024.

SHALEV-SHWARTZ, S.; BEN-DAVID, S. Understanding Machine Learning: From Theory to Algorithms. *Cambridge University Press*, 2014.

SOUSA, A. B. M. Trabalho de Conclusão de Curso de Ciência da Computação, Validação externa dos agrupamentos de eletrofácies com aprendizado não supervisionado utilizando litologias interpretadas em poços de petróleo. 2024.

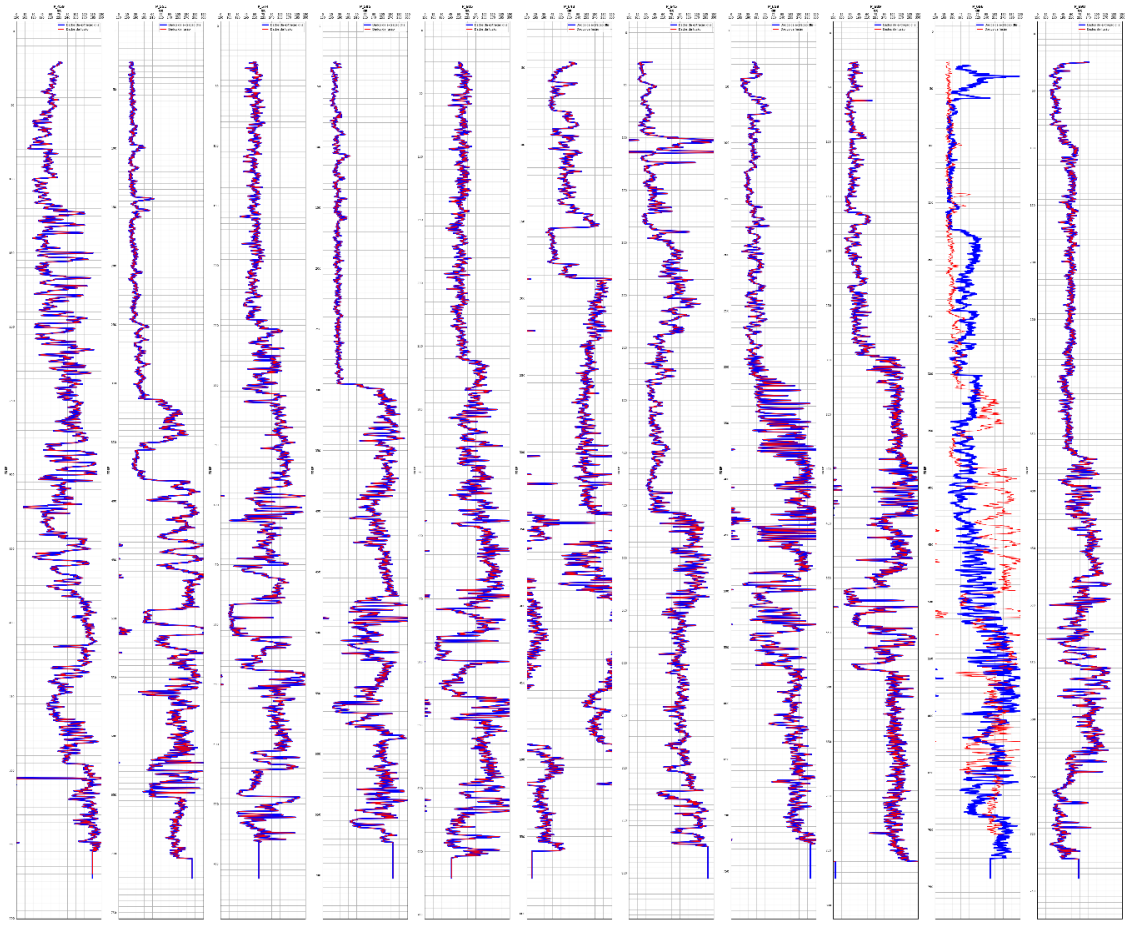
TCHAKOUCHT, T. et al. Random forest with feature selection and K-fold cross validation for predicting the electrical and thermal efficiencies of air based photovoltaic-thermal systems, 2024.

THOMAS, J. et al. Fundamentos da Engenharia de Petróleo. [S.l.]: Editora Interciência, 2001.

8. Outras atividades

Durante o ciclo de iniciação científica de 2023-2024, além da execução do plano de trabalho, houve outras atividades relacionadas ao estudo. Uma delas foi a verificação da fusão dos dados AGP com os dados DLIS, realizada com base na profundidade. A fim de se certificar que o encaixe das informações estava correto, foram criados gráficos para visualizar a curva GR nos dados obtidos do DLIS e da fusão do DLIS com o AGP. A figura 6 apresenta os 11 gráficos de cada poço criados para essa finalidade, onde a linha azul representa os dados extraídos dos DLIS, e a vermelha os dados da fusão. Pode-se observar que houve a sobreposição das linhas em 10 dos 11 gráficos, o único poço que apresentou problemas foi o P_696, segundo da direita para esquerda. Desse modo, por meio dessa verificação, foi possível detectar inconsistências no processamento da fusão, das quais foram posteriormente corrigidas.

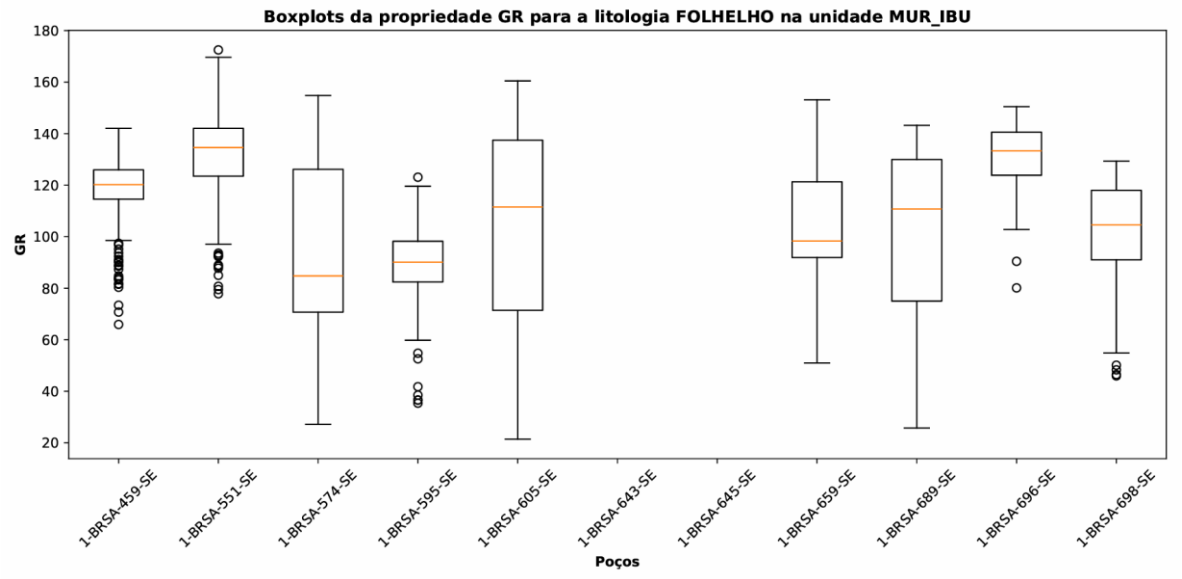
Figura 6 – Gráficos para a verificação da fusão DLIS com o AGP



Fonte: Autor

Ademais, foram criados gráficos do tipo box plot (diagrama de caixas) para fins de análise das propriedades dos tipos litológicos em determinada unidade geológica. A figura 7 mostra os box plots da curva de perfil GR das amostras de Folhelho encontradas na unidade MUR_IBU.

Figura 7 – Box plots da propriedade GR para a litologia Folhelho na unidade MUR_IBU



Fonte: Autor