UNIVERSIDADE FEDERAL DE OURO PRETO - UFOP
ESCOLA DE MINAS
DEPARTAMENTO DE ENGENHARIA DE PRODUÇÃO

GABRIEL SANGLARD SENRA

UTILIZAÇÃO DO ALGORITMO RANDOM FOREST PARA GERAÇÃO DE PERFIS SÔNICOS SINTÉTICOS POR MEIO DE REGISTROS CONVENCIONAIS

OURO PRETO - MG
202

Gabriel Sanglard Senra

Utilização do algoritmo Random Forest para geração de perfis sônicos sintéticos por meio de registros convencionais

Monografia apresentada ao Curso de Graduação em Engenharia de Produção da Universidade Federal de Ouro Preto como requisito parcial para a obtenção do título de Engenheiro de Produção.

Professor Orientador: Prof. Dr. Joney Justo da Silva

OURO PRETO

2024

SISBIN - SISTEMA DE BIBLIOTECAS E INFORMAÇÃO

S478u Senra, Gabriel Sanglard.
Utilização do algoritmo Random Forest para geração de perfis sônicos sintéticos por meio de registros convencionais. [manuscrito] / Gabriel Sanglard Senra. - 2024.
54 f.: il.: color., gráf., tab., mapa.

Orientador: Prof. Dr. Joney Silva.
Monografia (Bacharelado). Universidade Federal de Ouro Preto. Escola de Minas. Graduação em Engenharia de Produção .

1. Registros sonoros. 2. Prospecção - Métodos geofísicos. 3. Bacias (Geologia). 4. Reservatórios subterrâneos. I. Silva, Joney. II. Universidade Federal de Ouro Preto. III. Título.

CDU 658.5

Bibliotecário(a) Responsável: Cristiane Maria da Silva - CRB6-3046

MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DE OURO PRETO
REITORIA
ESCOLA DE MINAS
DEPARTAMENTO DE ENGENHARIA DE PRODUCAO, ADMINISTRACAO E ECON

FOLHA DE APROVAÇÃO

Gabriel Sanglard Senra

Utilização do Algoritmo Random Forest para Geração de Perfis Sônicos Sintéticos por meio de Registros Convencionais

Aprovada em 01 de outubro de 2024

Monografia apresentada ao Curso de Engenharia de Produção da Universidade Federal de Ouro Preto como requisito parcial para obtenção do título de Engenheiro de Produção

Aprovada em 01 de outubro de 2024

Membros da banca

Dr. Joney Justo da Silva - Orientador (Universidade Federal de Ouro Preto)
Dr. Magno Silvério Campos - (Universidade Federal de Ouro Preto)
Dr. Jadson Castro Gertrudes - (Universidade Federal de Ouro Preto)

Dr. Joney Justo da Silva, orientador do trabalho, aprovou a versão final e autorizou seu depósito na Biblioteca Digital de Trabalhos de Conclusão de Curso da UFOP em 15/10/2024

Documento assinado eletronicamente por **Joney Justo da Silva**, **PROFESSOR DE MAGISTERIO SUPERIOR**, em 15/10/2024, às 14:51, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

A autenticidade deste documento pode ser conferida no site http://sei.ufop.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0 , informando o código verificador **0795575** e o código CRC **3DEC48D0**.

R. Diogo de Vasconcelos, 122, - Bairro Pilar Ouro Preto/MG, CEP 35402-163
Telefone: 3135591540 - www.ufop.br

Dedico este trabalho aos meus pais, irmãos, amigos e professores, cuja presença, apoio e incentivo foram essenciais ao longo desta jornada.

AGRADECIMENTO

Gostaria de expressar minha profunda gratidão aos meus pais, Pedro de Alcantara Senra do Oliveira Filho e Priscilla Drummond Sanglard, pelo seu inestimável amor, carinho e apoio constante ao longo da minha jornada. Eles foram meu porto seguro e me apoiaram incansavelmente em cada etapa. Sou igualmente grato pelo companheirismo e amor dos meus irmãos, Pedro de Alcantara Senra de Oliveira Neto e Jade Sanglard Senra, cujo apoio foi essencial para me manter firme na busca dos meus sonhos.

Meu agradecimento se estende aos professores Cristiano Luiz Turbino de Franca e Silva e Magno Silvério Campos, cuja dedicação e entusiasmo foram pilares na minha formação acadêmica. Eles não apenas compartilharam conhecimento, mas também serviram como inspiração e exemplos a serem seguidos.

Um agradecimento especial aos meus amigos, Lucas Santos de Sá e Felipe Péret Sasdelli, cuja amizade e suporte foram fundamentais para alcançar este marco em minha vida. A união e o apoio deles foram indispensáveis.

Estou imensamente grato por tudo que recebi e almejo, sinceramente, poder retribuir essa generosidade e apoio no futuro.

R E S U M O

Este estudo explora o uso de aprendizado de máquina, focando no modelo Random Forest, para inferir o perfil sônico a partir de registros geofísicos convencionais, como ILD, RHOB, GR e NPHI, escolhidos por sua relevância na caracterização geológica e forte correlação com as velocidades sônicas. Os resultados mostram que o Random Forest supera a Equação de Faust, referência empírica tradicional, com maior precisão e coeficiente de determinação (R^2), capturando melhor as complexidades geológicas do campo Lagoa Parda Sul. O estudo sugere melhorias, como otimização de hiperparâmetros, segmentação geológica e testes com outros algoritmos, como redes neurais e gradiente boosting.

Palavras chaves: Inferência do Perfil Sônico, Random Forest, Aprendizado de Máquina em Geofísica, Campo Lagoa Parda Sul, Bacia do Espírito Santo, Modelagem Geológica, Inteligência Artificial na Análise de Reservatórios.

A B S T R A C T

This study explores the use of machine learning, focusing on the Random Forest model, to infer the sonic log from conventional geophysical logs such as ILD, RHOB, GR, and NPHI, selected for their relevance in geological characterization and strong correlation with sonic velocities. The results show that Random Forest outperforms the Faust Equation, a traditional empirical reference, with higher accuracy and coefficient of determination (R^2), better capturing the geological complexities of the Lagoa Parda Sul field. The study suggests improvements, such as hyperparameter optimization, geological segmentation, and testing with other algorithms like neural networks and gradient boosting.

Keywords: Synthetic log; Sonic Log Inference, Random Forest, Machine Learning in Geophysics, Lagoa Parda Sul Field, Espírito Santo Basin, Geological Modeling, Artificial Intelligence in Reservoir Analysis.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

CAPÍTULO 1
INTRODUÇÃO

O registro de poços é uma ferramenta essencial para a caracterização de reservatórios e a avaliação de recursos de hidrocarbonetos. Ele desempenha um papel fundamental na identificação das propriedades petrofísicas e geomecânicas das rochas, permitindo um entendimento detalhado das formações subterrâneas (Gamal, 2021). Com o uso de perfis acústicos e sísmicos, é possível derivar parâmetros críticos para a engenharia de reservatórios e operações de perfuração.

Em geral, as ferramentas padrão de registro de poço medem o raio gama, a densidade da formação, as resistividades em diferentes penetrações e a porosidade. Essas são empregadas em avaliações gerais de formações, fornecendo informações petrofísicas e litológicas. Além disso, como abordagem adicional, o registro acústico é realizado para identificar características petrofísicas e mecânicas avançadas. Os registros sonoros de compressão e cisalhamento são essenciais para derivar não apenas propriedades mecânicas, mas também sua aplicação na análise sísmica quantitativa. (Ameen, 2009; Rasouli, 2011)

Nos últimos anos, aplicações avançadas que utilizam registros sonoros têm sido empregadas na avaliação de sistemas de hidratos de gás, os quais são esperados para fornecer implicações geológicas adicionais na exploração e exploração subsequente. (Saumya, 2019)

Várias variáveis geofísicas e geológicas, como densidade, rigidez, umidade, porosidade, composição mineral, temperatura e pressão, influenciam a velocidade do som nas camadas superficiais da Terra. Ao considerar essas variáveis e suas interações, é possível fazer previsões mais precisas do perfil sônico, o que é crucial para entender a estrutura e as propriedades do subsolo da Terra, bem como para aplicações na exploração de recursos naturais e engenharia geotécnica.

Recentemente, o uso de algoritmos de aprendizado de máquina, como o Random Forest, tem ganhado destaque na previsão em tempo real dos perfis sônicos,

oferecendo uma alternativa eficiente e precisa em comparação aos métodos tradicionais de correlação (Gamal, 2021). Estes algoritmos permitem uma melhor predição de atributos em grandes volumes de dados geológicos, tornando o processo de interpretação mais rápido e preciso. Além disso, o uso dessas técnicas está revolucionando a exploração de recursos naturais, proporcionando avanços significativos no monitoramento ambiental e na exploração de hidrocarbonetos (Lin, 2016).

Nesse contexto, algoritmos artificiais como aprendizado de máquina, redes neurais, algoritmos genéticos e outros métodos de inteligência artificial emergiram como ferramentas poderosas para inferência de atributos em perfis estratigráficos. Esses algoritmos são capazes de extrair informações valiosas, identificar padrões ocultos e fazer previsões precisas com base em grandes volumes de dados estratigráficos.

É possível perceber que algoritmos artificiais estão revolucionando a interpretação de perfis estratigráficos, melhorando eficiência, precisão e consistência nas análises geológicas. Ao automatizar e otimizar os processos de interpretação, essas técnicas proporcionam uma compreensão mais profunda das formações geológicas e têm aplicações significativas em campos como exploração de recursos naturais, geotecnia, monitoramento ambiental e estudos geotécnicos, contribuindo para o contínuo avanço do conhecimento geológico e geofísico.

1.1 - Localização

A área de exploração que será abordada é conhecida como Campo de Lagoa Parda Sul, com uma extensão de desenvolvimento de aproximadamente 1,73 km², encontra-se situada na parte terrestre da Bacia do Espírito Santo. Este campo está localizado no município de Linhares, a uma distância aproximada de 118 km a nordeste da cidade de Vitória, que é a capital do estado do Espírito Santo. A Figura 1 apresenta a localização da área com destaque ao campo de produção estudado e aos poços petrolíferos utilizados neste trabalho.

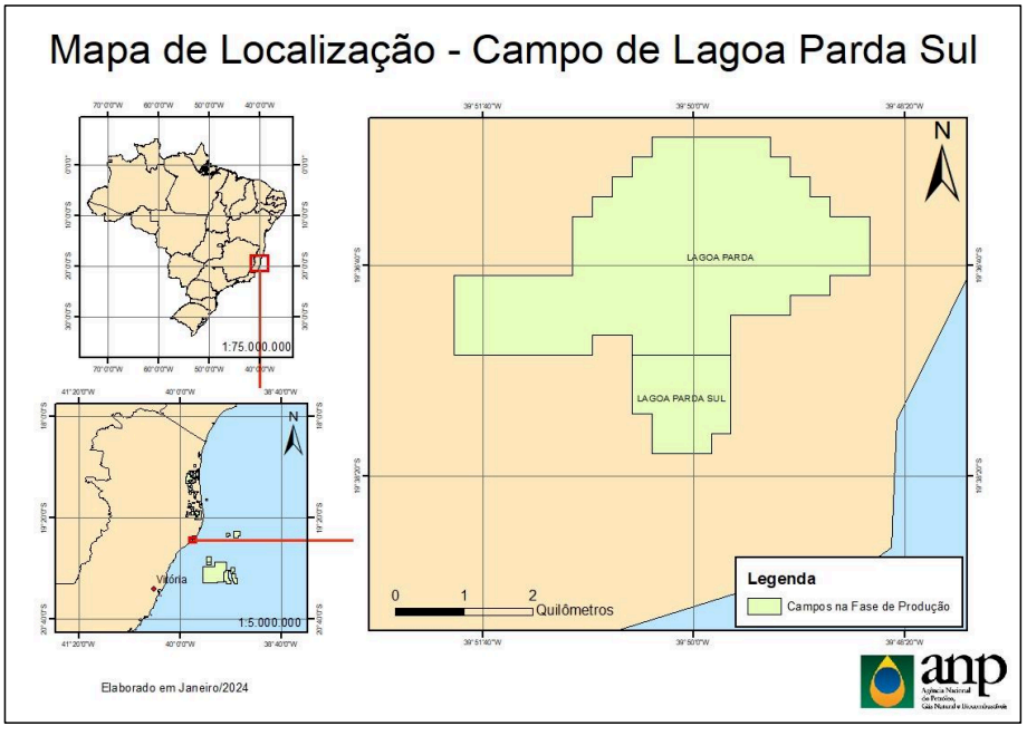

Figura 1 - Localização da área de estudo com destaque ao campo de produção Lagoa Parda e Lagoa Parda Sul. Fonte: Relatório do Plano de Desenvolvimento ANP, 2024

1.2 - Objetivo

O principal objetivo deste trabalho é inferir o tempo de trânsito compressional (perfil sônico) através de algoritmos de aprendizado de máquina supervisionados em modelos de regressão, utilizando-se dados de perfilagem geofísica. Para tanto, são tidos como objetivos específicos:

- Aplicação do modelo de aprendizado de máquina nos perfis estratigráficos do campo Lagoa Parda Sul
- Graficar e comparar o valor predito do perfil sônico em relação ao seu registro original.
- Avaliar o desempenho do modelo de inteligência artificial em relação a métodos convencionais de dedução do perfil sônico.

1.3 - Justificativa

A possibilidade de inferir o registro sônico em poços que não possuem essa informação é de grande relevância para a indústria de petróleo e gás, pois permite economizar recursos financeiros significativos e possibilita análises em poços onde a coleta direta de dados enfrenta restrições técnicas. A utilização de métodos estatísticos e algoritmos de aprendizado de máquina para estimar registros sônicos a partir de outros perfis geofísicos disponíveis proporciona uma alternativa eficiente e econômica. Isso não só reduz os custos operacionais associados à aquisição de dados, mas também amplia a capacidade de avaliação de formações geológicas, viabilizando a exploração e desenvolvimento de reservatórios em situações desafiadoras onde a obtenção direta de dados sônicos é impraticável (Yu, Yanxiang & Xu, 2021).

Os registros de tempo de trânsito sísmico englobam dados geomecânicos essenciais para caracterizar o subsolo nas proximidades do poço. Frequentemente, os registros sísmicos são indispensáveis na finalização do processo de amarração sísmica do poço ou na previsão de propriedades geomecânicas. Em situações em que um poço ou um intervalo específico não possui registros sísmicos, um método comumente empregado envolve a criação de registros sintéticos com base em poços vizinhos que possuem dados sísmicos. Esse processo é conhecido como síntese de registros sísmicos ou geração de registros sísmicos pseudos.

CAPÍTULO 2
CONTEXTO GEOLÓGICO

A região da Lagoa Parda, situada em Linhares, Espírito Santo, Brasil, integra a Bacia do Espírito Santo, uma bacia sedimentar significativa pela sua riqueza em recursos petrolíferos e hídricos. Segundo Valadão e Lima (2017), esta área é predominantemente composta por sedimentos da Bacia do Espírito Santo, que inclui uma sequência de arenitos, argilitos e folhelhos, depositados durante o período Cretáceo. Almeida (1977) destaca que a atividade tectônica na região influenciou a formação de várias falhas e dobras, contribuindo para a complexidade geológica local. Além disso, Schobbenhaus e Brito Neves (2003) indicam que os recursos hídricos subterrâneos são uma característica importante, com a presença de aquíferos significativos que sustentam a economia agrícola e a indústria local.

2.1 - Estratigrafia

Compreender a estratigrafia e a tectônica da área é essencial para a interpretação dos perfis sônicos dos poços de petróleo, permitindo uma avaliação precisa dos recursos subterrâneos.

A coluna estratigráfica da Bacia do Espírito Santo referente a Figura 2 apresenta uma linha do tempo geológica que ilustra as diferentes camadas de rochas e os eventos que ocorreram na região ao longo dos milhões de anos. No início, encontramos o embasamento Pré-Cambriano, que serve como base para todas as outras camadas. Sobre essa base, durante o período rifte, começaram a se formar sedimentos que criaram a Formação Cricaré, composta por conglomerados, arenitos, folhelhos lacustres e coquinas. As falhas tectônicas ajudaram a criar estruturas que influenciaram onde esses sedimentos se acumularam.

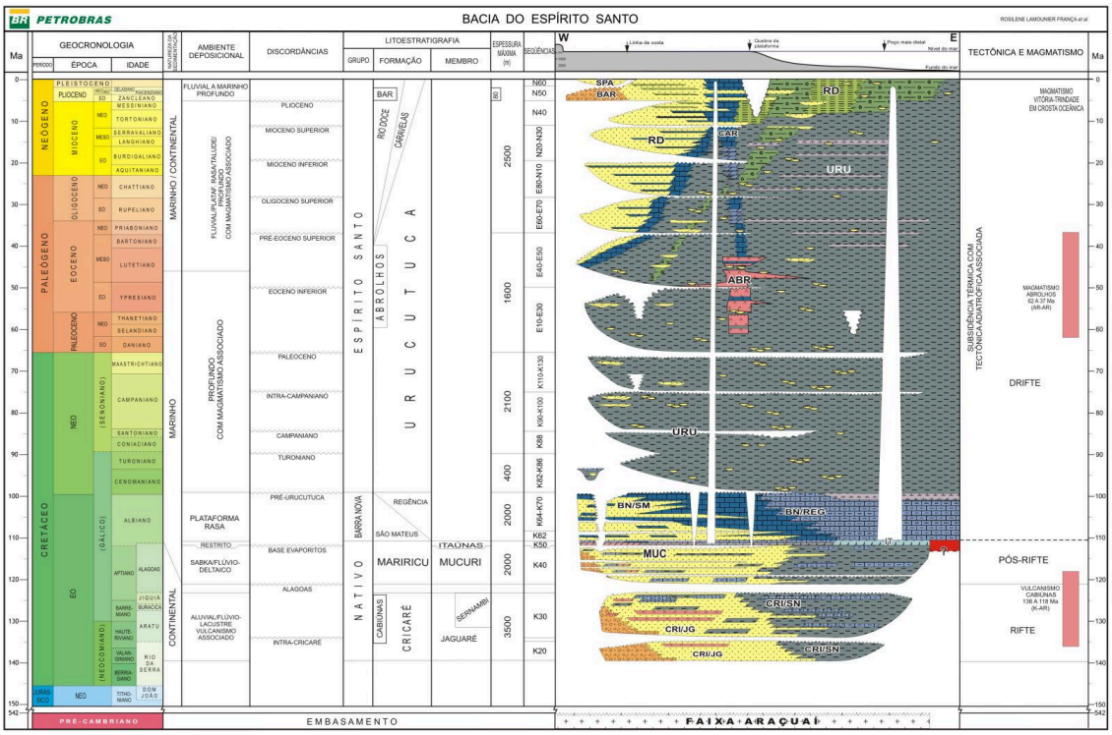

Figura 2: Diagrama estratigráfico da Bacia do Espírito Santo (FRANÇA, 2007).

O Sumário Geológico e Setores em Oferta da ANP (2021) destaca que o Pós-Rifte da Bacia do Espírito Santo (rifte-sag) é representado por pacotes de sedimentos siliciclásticos do Membro Mucuri e evaporíticos do Membro Itaúnas da Formação Mariricu, registro das primeiras incursões marinhas na bacia.

Na fase pós-rifte, começou a deposição de sedimentos finos em ambientes marinhos, formando os membros Mucuri e Itaúnas da Formação Mariricu. Estes sedimentos indicam que o mar começou a invadir a bacia, trazendo novos tipos de depósitos que podem armazenar petróleo e gás natural. A parte superior desta camada está em continuidade com o Grupo Barra Nova na área leste da bacia, mas apresenta uma descontinuidade em relação à Formação Urucutuca na área oeste.

A fase drifte, iniciada no Albiano e continuando até hoje, inclui o Grupo Barra Nova, com as formações São Mateus e Regência, compostas por arenitos e sedimentos carbonáticos. O Grupo Espírito Santo inclui as formações Urucutuca, Caravelas e Rio Doce. A Formação Urucutuca é especialmente importante devido aos

seus folhelhos e arenitos turbidíticos, que são cruciais para a exploração de petróleo e gás.

Em suma, é possível perceber que a coluna estratigráfica da Bacia do Espírito Santo nos conta a história das formações geológicas que compõem a bacia e como os eventos tectônicos criaram condições favoráveis para a formação de reservatórios de petróleo e gás.

2.2 - Tectônica na Bacia do Espírito Santo

A tectônica na Bacia do Espírito Santo teve um papel essencial na formação da estrutura e na distribuição dos sedimentos, influenciando diretamente o potencial de exploração de petróleo e gás. A história tectônica da bacia pode ser dividida em três fases principais: rifte, pós-rifte e drifte.

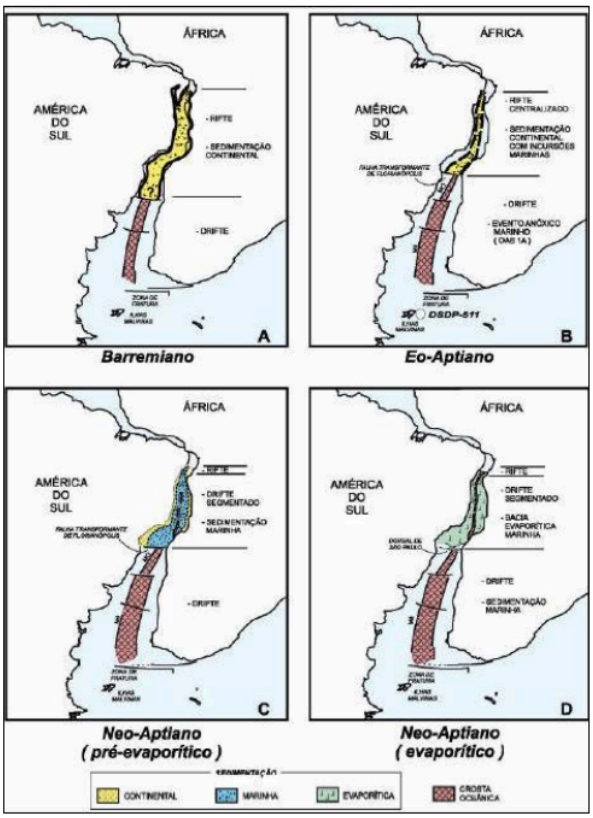

Figura 3: Reconstrução paleogeográfica do Atlântico Sul durante as fases rifte e transicional. MOHRIAK, 2003

Na fase rifte, que ocorreu há milhões de anos, a crosta terrestre se esticou e quebrou, formando grandes falhas. Essas falhas criaram vales profundos e montanhas baixas, onde sedimentos começaram a se acumular. Foi durante essa fase que as formações iniciais de rochas vulcânicas e sedimentares se estabeleceram (França, 2007).

Seguindo a fase rifte, a fase pós-rifte trouxe um período de maior estabilidade, com menos atividade tectônica. Durante esse tempo, a área foi gradualmente coberta por sedimentos marinhos, formando camadas importantes que são hoje vitais para a exploração de recursos naturais (França, 2007).

A fase drifte, que começou mais recentemente, viu a bacia se aprofundar ainda mais, com uma sedimentação contínua em ambientes marinhos profundos. Movimentos de sal nas camadas subterrâneas criaram estruturas complexas, que são cruciais para a formação de reservatórios de petróleo e gás. Além disso, a atividade vulcânica durante este período, especialmente os derrames de lava, moldou ainda mais a bacia (Agência Nacional do Petróleo, Gás Natural e Biocombustíveis, 2021).

2.3 – Hidrogeologia na Bacia do Espírito Santo

Na Bacia do Espírito Santo, os principais aquíferos estão localizados nas formações sedimentares que datam desde o período Cretáceo até o Quaternário. Estas formações incluem os arenitos da Formação São Mateus e os depósitos flúvio-deltaicos do Membro Mucuri da Formação Mariricu. A qualidade e a porosidade dessas rochas determinam a capacidade de armazenamento e a transmissividade dos aquíferos, influenciando a resposta sônica durante as perfurações. Segundo Valadão e Lima (2017), os aquíferos mais significativos são aqueles que apresentam alta porosidade e permeabilidade, permitindo um fluxo de água mais eficiente.

A presença de água nos aquíferos pode causar variações na velocidade das ondas sônicas, que são usadas para determinar a composição e a estrutura das rochas subterrâneas. Quando a água está presente em grandes volumes, ela tende a diminuir a velocidade das ondas sônicas devido à sua menor densidade e compressibilidade em

comparação com os hidrocarbonetos. Isso significa que, durante a perfilagem sônica, a interpretação dos dados deve levar em conta a saturação de água nos aquíferos para evitar erros na estimativa da localização e da quantidade de petróleo e gás (Schobbenhaus & Brito Neves, 2003).

A compreensão detalhada das propriedades dos aquíferos e sua influência nos perfis sônicos permite uma avaliação mais precisa dos reservatórios de petróleo e gás, otimizando a exploração e minimizando riscos. A integração de dados hidrogeológicos com perfis sônicos é, portanto, essencial para a prospecção bem-sucedida na região.

CAPÍTULO 3

PERFILAGEM GEOFÍSICA

Perfilagem geofísica é uma técnica crucial na exploração de petróleo e gás, envolvendo a medição das propriedades físicas das formações geológicas através de ferramentas inseridas em poços de perfuração. Essas ferramentas registram dados em tempo real sobre a resistividade elétrica, densidade, velocidade sônica, porosidade e outros parâmetros das rochas e fluidos subterrâneos. Esses dados são fundamentais para a caracterização dos reservatórios, permitindo a identificação de zonas produtivas e a avaliação da qualidade dos reservatórios. Através da interpretação dos perfis geofísicos, é possível determinar a saturação de água e hidrocarbonetos, a permeabilidade das formações e a distribuição dos poros, informações essenciais para a tomada de decisões na perfuração e produção de petróleo e gás (Assaad, 2009; Azeem, 2016; Johnson, 1962).

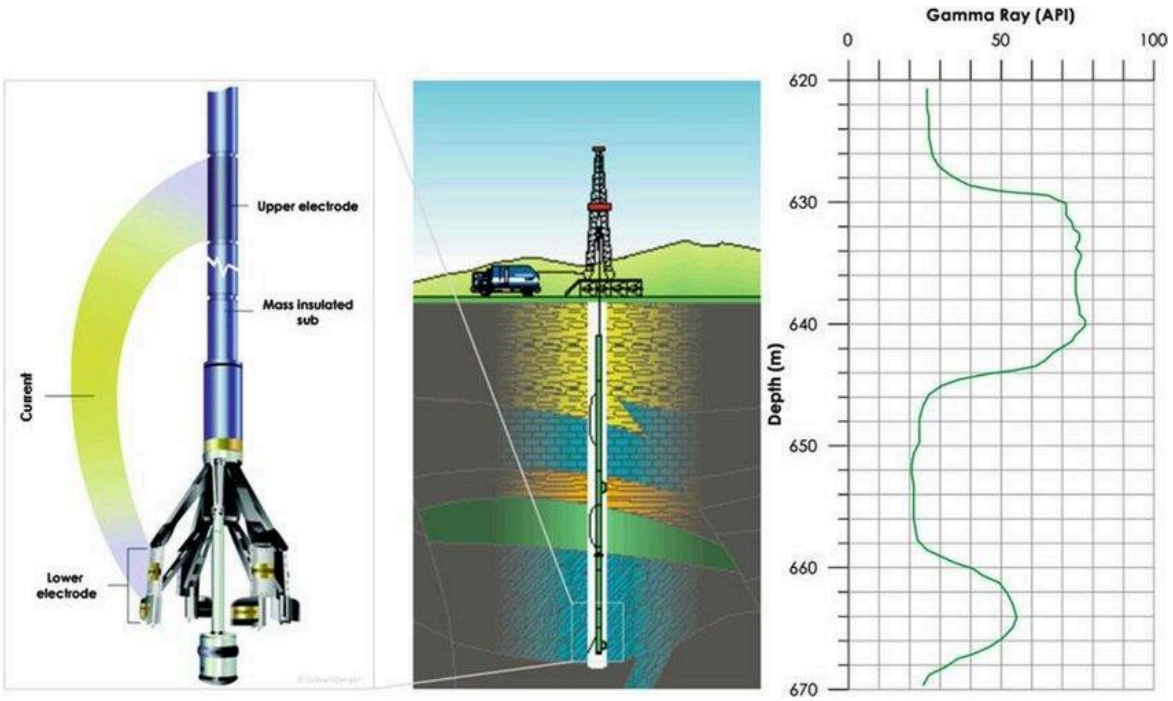

Figura 4: Exemplo de uma ferramenta de perfilagem. A operação de perfilagem de poços mostrando o caminhão de perfilagem, o cabo de perfilagem estendido na plataforma de perfuração.

A Figura 4 ilustra o processo de perfilagem geofísica em poços, essencial para a caracterização das formações geológicas e a identificação de zonas produtivas de petróleo e gás. À esquerda, um diagrama detalhado mostra uma ferramenta de perfilagem equipada com eletrodos superiores e inferiores, utilizada para medir diferentes propriedades geofísicas das formações geológicas circundantes. No centro, a configuração no poço de petróleo é representada, destacando a descida da ferramenta a partir da superfície, onde uma torre de perfuração está instalada. À direita, o gráfico de perfil de raios gama exibe a variação da radioatividade natural das formações em função da profundidade. Esta representação visual é fundamental para compreender como os dados são coletados e utilizados na análise geológica e na tomada de decisões na indústria de petróleo e gás.

3.1 Mnemônicos na Perfilagem Geofísica de Poços

A perfilagem geofísica é uma técnica essencial na caracterização de formações subterrâneas, permitindo a análise detalhada das propriedades físicas das rochas e dos fluidos presentes. Para alcançar essa compreensão, uma variedade de registros geofísicos, conhecidos como mnemônicos, são utilizados. Esses mnemônicos representam diferentes medições físicas que, em conjunto, fornecem uma visão abrangente das características da formação.

Entre os mnemônicos mais amplamente utilizados está o Gamma Ray (GR), que "auxilia na identificação de litologias e na estimativa da argilosidade" (Asquith & Krygowski, 2004). O Spontaneous Potential (SP) mede "diferenças de potencial elétrico para identificar zonas permeáveis e calcular a resistividade da água de formação" (Rider & Kennedy, 2011). O Bulk Density (RHOB) fornece informações sobre a "densidade aparente da formação, crucial para a determinação da porosidade" (Ellis & Singer, 2007).

Além disso, o Neutron Porosity (NPHI) e o Sonic Travel Time (DT) são indispensáveis para a avaliação da porosidade, sendo o primeiro "sensível à presença de hidrogênio" (Asquith & Krygowski, 2004) e o segundo baseado na "velocidade de

propagação de ondas sônicas" (Schlumberger, 1989). O Induction Log Deep (ILD) mede a "resistividade em profundidade, fundamental para a avaliação da saturação de hidrocarbonetos" (Schlumberger, 1989), enquanto o Caliper (CALI) monitora o "diâmetro do poço, garantindo a qualidade e precisão dos outros registros" (Rider & Kennedy, 2011).

Esses mnemônicos, ao serem interpretados de forma integrada, oferecem uma base sólida para a construção de modelos geológicos e petrofísicos, sendo fundamentais na tomada de decisões em atividades de exploração e produção de recursos naturais.

3.1.1 Perfil de Raio Gama

O registro de *Gamma Ray* (GR) mede a intensidade da radiação gama natural emitida pelas formações geológicas, permitindo a identificação de litologias e a distinção entre folhelhos e arenitos. Os xistos tendem a exibir altos valores de GR devido à presença de argilas ricas em elementos radioativos, como urânio, tório e potássio. Este registro é particularmente útil para a correlação estratigráfica entre poços e para a determinação da argilosidade das formações.

O índice de argilosidade (I_{cl}) é calculado pela fórmula:

$$I_{cl} = \frac{GR_{log} - GR_{min}}{GR_{max} - GR_{min}} \qquad \text{Equação (3.1.1)}$$

Onde:

- GR_{log} é o valor de GR registrado
- GR_{min} é o valor mínimo de GR (geralmente para arenitos limpos).
- GR_{max} é o valor máximo de GR (geralmente para xistos).

A aplicação do GR é extensa, incluindo a correção de perfis de porosidade e resistividade para efeitos de argilosidade. "O registro GR é amplamente utilizado para

correlacionar camadas sedimentares entre diferentes poços, auxiliando na interpretação estratigráfica" (Asquith & Krygowski, 2004).

3.1.2 Perfil de Resistividade

O perfil de resistividade tem como objetivo medir a capacidade das formações geológicas em resistir ao fluxo de corrente elétrica, sendo amplamente utilizado na caracterização de reservatórios. Através dessa medição, torna-se possível identificar a presença e distribuição de fluidos, como água e hidrocarbonetos, no interior das rochas. Essa informação é fundamental para a avaliação da qualidade do reservatório e para a estimativa de sua capacidade produtiva. Diferentes tipos de perfis de resistividade são utilizados dependendo da profundidade e do raio de investigação necessário, cada um com características técnicas específicas que permitem maior precisão na interpretação das propriedades das formações.

3.1.2.1 Perfil de Resistividade Induzida

O *Induction Log Deep* (ILD) mede a resistividade elétrica das formações em profundidade. Este registro é fundamental para identificar zonas de hidrocarbonetos, especialmente em formações de baixa porosidade ou altamente resistivas.

A saturação de água (S_w) pode ser estimada pela equação de Archie:

$$S_w = \left(\frac{a \times R_w}{\phi^m \times R_t} \right)^{1/n} \qquad \text{Equação (3.1.2.1)}$$

Onde:

- S_w é a saturação de água;
- $a,\ m\ e\ n$ são constantes empíricas;
- R_w é a resistividade da água de formação;
- R_t é a resistividade medida da formação;

O ILD é amplamente utilizado para determinar a saturação de hidrocarbonetos e para a avaliação da qualidade dos reservatórios. "Este registro é crucial para a avaliação da saturação de fluidos em formações profundas e resistivas" (Schlumberger, 1989).

3.1.2.2 Perfil de Resistividade de Curto Alcance (SFL)

O *Spherically Focused Log* (SFL) mede a resistividade em regiões próximas à parede do poço. Ele é especialmente útil para identificar camadas finas ou permeáveis, onde outros perfis de resistividade podem não ser tão sensíveis. O SFL complementa o ILD ao fornecer informações adicionais sobre a distribuição de fluidos mais próxima do poço.

3.1.2.3 Perfil de Resistividade Microfocalizada (MSFL)

O *Micro Spherically Focused Log* (MSFL) é projetado para medir a resistividade em uma área muito próxima à parede do poço, geralmente na zona invadida por fluidos de perfuração. Ele ajuda a diferenciar entre as formações verdadeiras e as zonas alteradas pelo processo de perfuração, auxiliando na interpretação da qualidade da formação e na identificação de hidrocarbonetos presentes nas camadas mais superficiais.

3.1.3 Perfis Neutrônicos

O *Neutron Porosity* (NPHI) mede a porosidade da formação com base na absorção de neutrons rápidos. Este registro é sensível à presença de hidrogênio, sendo assim útil na determinação da porosidade efetiva em formações contendo água ou hidrocarbonetos.

A porosidade neutrons (ϕ_N) é dada por:

$$\phi_N = \frac{C_{np}}{\rho_{mat}} - \frac{C_f}{\rho_f} \qquad \text{Equação (3.1.3)}$$

Onde:

- C_{np} e C_f são constantes de correção de porosidade;
- ρ_{mat} e ρ_f são as densidades da matriz e do fluido, respectivamente;

O NPHI é particularmente valioso para a identificação de zonas de gás, onde a porosidade medida pode diferir significativamente da porosidade real devido à presença de gás. "O NPHI é frequentemente utilizado em conjunto com o RHOB para diferenciar entre gás, óleo e água na formação" (Asquith & Krygowski, 2004).

3.1.4 Perfil de Densidade

O registro de *Bulk Density* (RHOB) mede a densidade aparente da formação, incluindo a densidade dos grãos minerais e dos fluidos nos poros. Este registro é fundamental para a determinação da porosidade da formação e, quando combinado com o registro de *Neutron Porosity* (NPHI), permite a diferenciação entre fluidos como gás, óleo e água.

A porosidade (ϕ) pode ser calculada pela fórmula:

$$\phi = \frac{\rho_m - \rho_b}{\rho_m - \rho_f} \qquad \text{Equação (3.1.4)}$$

Onde:

- ρ_m é a densidade da matriz mineral;
- ρ_b é a densidade aparente da formação;
- ρ_f é a densidade do fluido nos poros;

O RHOB é amplamente utilizado para determinar a composição mineralógica das formações e a presença de diferentes tipos de fluidos. "A densidade aparente medida pode ser utilizada para calcular a porosidade da formação e identificar litologias distintas" (Ellis & Singer, 2007).

3.1.5 Perfil Sônico

O Sonic Travel Time (DT) mede o tempo que uma onda sônica leva para atravessar uma determinada distância na formação. Este registro é crucial para calcular a porosidade da formação utilizando a equação de Wyllie:

$$\phi = \frac{DT - DT_m}{DT_f - DT_m} \qquad \text{Equação (3.1.5)}$$

Onde:

- DT é o tempo de viagem medido;
- DT_m é o tempo de viagem da matriz sólida;
- DT_f é o tempo de viagem do fluido de poros;

Além de calcular a porosidade, o registro DT é utilizado para avaliar a compactação da rocha e a integridade mecânica das formações. "O DT é essencial na modelagem de perfil sônico preditivo e na determinação de parâmetros petrofísicos críticos" (Schlumberger, 1989).

3.1.6 Perfil Caliper

O Caliper (CALI) é um registro que mede o diâmetro interno do poço, fornecendo informações cruciais sobre a rugosidade das paredes do poço e possíveis cavitações ou desmoronamentos.

O registro de caliper pode ser utilizado para calcular o volume de lodo necessário para preencher o poço e para detectar zonas de fratura ou formação

instável. "O CALI é essencial para garantir a precisão dos outros registros de perfilagem, especialmente em poços com diâmetros irregulares" (Rider & Kennedy, 2011).

CAPÍTULO 4
INTELIGÊNCIA ARTIFICIAL

A inteligência artificial é definida como a capacidade dos sistemas computacionais de realizar tarefas que normalmente exigiriam inteligência humana. Conforme descrito por McCarthy (1955), a IA refere-se à criação de programas de computador que podem realizar tarefas que, quando realizadas por seres humanos, requerem inteligência.

4.1 Definição de Aprendizado de Máquina

O aprendizado de máquina é uma subárea da inteligência artificial (IA) que se concentra no desenvolvimento de algoritmos que permitem aos computadores aprenderem a partir de dados e melhorar seu desempenho em tarefas específicas sem serem explicitamente programados para isso. Segundo Mitchell (1997), "um programa de computador aprende com a experiência E em relação a alguma classe de tarefas T e medida de desempenho P, se seu desempenho em tarefas em T, medido por P, melhora com a experiência E". De forma similar, Domingos (2012) define aprendizado de máquina como "a ciência de fazer computadores aprenderem a partir de dados, sem serem explicitamente programados".

A essência do aprendizado de máquina está na capacidade dos algoritmos de identificar padrões nos dados e utilizar esses padrões para fazer previsões ou tomar decisões sobre novos conjuntos de dados não vistos anteriormente. Isso é alcançado através da aplicação de técnicas estatísticas e computacionais que permitem aos modelos aprenderem de forma automática e iterativa (Bishop, 2006).

Existem diversos tipos de aprendizado de máquina, incluindo aprendizado supervisionado, não supervisionado e por reforço. Cada tipo tem suas próprias características e é adequado para diferentes tipos de problemas.

4.1.1 Aprendizado Supervisionado

No aprendizado supervisionado, os algoritmos são treinados com um conjunto de dados rotulados, onde cada exemplo de entrada está associado a uma saída desejada conhecida. O objetivo é aprender uma função que mapeie as entradas para as saídas, permitindo fazer previsões precisas em novos dados. Russell e Norvig (2016) explicam que "no aprendizado supervisionado, a tarefa do aprendiz é prever o valor de uma variável de destino com base em um conjunto de variáveis de entrada". Este tipo de aprendizado é amplamente utilizado em tarefas como classificação e regressão.

Além do aprendizado supervisionado, existem também os modelos de aprendizado não supervisionado e por reforço, que não são utilizados neste contexto.

4.2 Modelos de regressão

Os modelos de regressão são fundamentais no campo do aprendizado de máquina, especialmente no contexto de previsão de valores contínuos com base em variáveis de entrada. Segundo Hastie (2009), "a regressão é uma técnica estatística usada para estudar a relação entre variáveis". No contexto dos algoritmos de aprendizado de máquina, os modelos de regressão são utilizados para estimar ou prever valores numéricos, facilitando análises preditivas e decisões informadas em uma variedade de aplicações práticas.

A avaliação da performance do modelo é crucial. O erro quadrático médio (MSE) mede a média dos quadrados dos erros entre as previsões e os valores observados. Um MSE menor indica um melhor ajuste do modelo aos dados. O coeficiente de determinação (R^2) indica a proporção da variância na variável dependente que é previsível a partir das variáveis independentes. Um R^2 mais próximo de 1 indica um melhor ajuste do modelo aos dados.

4.2.1 Aplicação de Algoritmos de Regressão

No contexto do aprendizado supervisionado, o objetivo é treinar um modelo para prever uma variável de interesse, como o tempo de trânsito sônico, utilizando um conjunto de dados rotulados. Os algoritmos de regressão são especialmente adequados para essa tarefa, pois são desenvolvidos para prever valores contínuos.

Os modelos de regressão linear têm sido amplamente utilizados para a inferência do perfil sônico. Por exemplo, Gunning e Glinsky (2007) utilizaram regressão linear múltipla para prever o tempo de trânsito sônico a partir de perfis de raios gama, densidade e porosidade neutrônica. Da mesma forma, Jain e Saraf (2019) aplicaram redes neurais artificiais para prever a porosidade de formações a partir de dados de perfilagem sônica e outros logs.

4.3 Modelo de Regressão Random Forest

O modelo de regressão Random Forest é um algoritmo de aprendizado de máquina que se enquadra na categoria de aprendizado supervisionado. Desenvolvido por Leo Breiman em 2001, o Random Forest é uma extensão dos métodos de árvores de decisão, onde múltiplas árvores de decisão são construídas durante o treinamento e suas previsões são combinadas para melhorar a precisão e a robustez do modelo, sendo utilizados no contexto da regressão, para prever valores contínuos.

O algoritmo Random Forest funciona construindo um conjunto de árvores de decisão, cada uma treinada com um subconjunto diferente dos dados de treinamento. Esse processo de construção é conhecido como "*bootstrap aggregating*", onde amostras são selecionadas aleatoriamente com reposição. Além disso, durante a construção de cada árvore, uma seleção aleatória de características é considerada para a divisão em cada nó, promovendo a diversidade entre as árvores.

Após a construção das árvores, a previsão final do modelo de regressão é obtida através da média das previsões de todas as árvores individuais. Isso ajuda a reduzir a

variância e a evitar o *overfitting*, um problema comum em modelos de árvore de decisão única.

A escolha do algoritmo Random Forest para a inferência do perfil sônico a partir de registros geofísicos convencionais na exploração de petróleo justifica-se por várias razões relacionadas à sua robustez, precisão e capacidade de lidar com dados complexos e não lineares. Além disso, o Random Forest pode fornecer insights adicionais sobre a importância das variáveis, facilitando a interpretação de quais registros têm maior peso na previsão do perfil sônico.

Em comparação com outros algoritmos de aprendizado supervisionado, como redes neurais ou gradiente boosting, o Random Forest oferece vantagens operacionais. Redes neurais, por exemplo, apesar do elevado potencial em termos de precisão preditiva, demandam mais tempo de treinamento e exigem maior conhecimento técnico para o ajuste dos hiperparâmetros, além de serem menos fáceis de interpretar. Diante dessas considerações, o Random Forest oferece um equilíbrio entre desempenho preditivo, facilidade de uso e capacidade de generalização, sendo a escolha adotada para a inferência de perfis sônicos neste cenário.

4.3.1 Fórmula do Modelo de Regressão Random Forest

Matematicamente, o Random Forest pode ser representado como um conjunto de de B árvores de decisão, onde cada árvore $Ti(x)$ é treinada com um subconjunto dos dados de treinamento. A previsão final $\hat{y}(x)$ para uma entrada x é calculada pela média das previsões das B árvores:

$$\hat{y}(x) = B1\sum i = 1BTi(x) \qquad \text{Equação (4.3.1)}$$

Onde, $\hat{y}(x)$ representa a previsão final, B é o número total de árvores na floresta, e $Ti(x)$ é a previsão da i-ésima árvore de decisão para a entrada x.

O processo de construção de cada árvore Ti no Random Forest envolve duas etapas principais. Primeiro, é realizada uma amostragem por bootstrapping, onde se selecionam aleatoriamente, com reposição, N amostras do conjunto de dados de treinamento para treinar cada árvore. Isso garante que cada árvore seja treinada com um subconjunto diferente dos dados, promovendo diversidade entre as árvores. Em seguida, para cada divisão em um nó da árvore, seleciona-se aleatoriamente um subconjunto de *features* m. A partir desse subconjunto, escolhe-se a melhor feature para realizar a divisão, o que ajuda a aumentar a variabilidade entre as árvores e a reduzir a correlação entre elas. Essas etapas combinadas garantem que o modelo final seja robusto e menos suscetível ao *overfitting*.

Para cada nó de uma árvore, a função de divisão é escolhida de forma a minimizar a soma das variâncias dentro dos nós resultantes. Se S é o conjunto de dados no nó atual, a divisão é realizada escolhendo a feature j e o ponto de divisão s que minimizam a expressão:

$$\left(\frac{1}{|S_1(j,s)|}\sum_{i\in S_1(j,s)}\left(y_i-\overline{y_{S_1}}\right)^2+\frac{1}{|S_2(j,s)|}\sum_{i\in S_2(j,s)}\left(y_i-\overline{y_{S_2}}\right)^2\right) \qquad \text{Equação (4.3.1)}$$

Onde:

- $S_1(j,s)$ e $S_2(j,s)$ são os subconjuntos de S criados pela divisão usando a *feature* j e o ponto de divisão s
- $\overline{y}S_1$ e $\overline{y}S_2$ são as médias das respostas nos subconjuntos S_1 e S_2, respectivamente.
- $|S_1(j,s)|$ e $|S_2(j,s)|$ representam o número de amostras em cada subconjunto.

Essa fórmula busca minimizar a soma das variâncias dentro dos nós resultantes, selecionando a melhor *feature* j e o ponto de divisão s que promovam essa minimização.

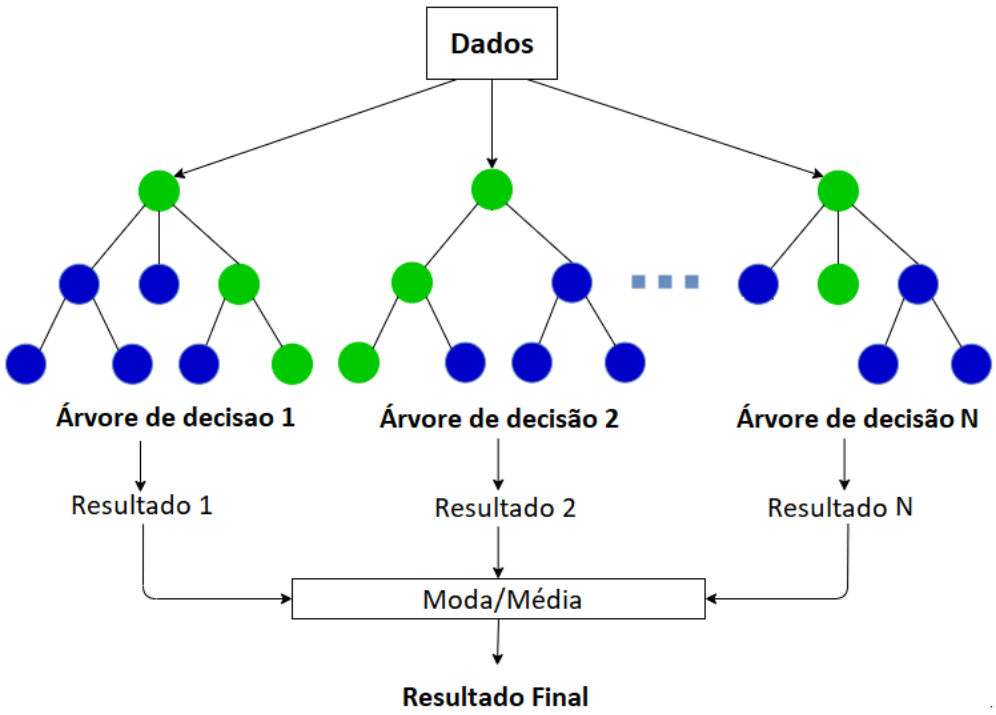

Figura 5: Representação visual do algoritmo de regressão Randon Forest.

4.3.2 Aplicação de Inteligência Artificial na Inferência do Perfil Sônico

A aplicação de inteligência artificial (IA) na inferência do perfil sônico tem se mostrado uma abordagem promissora para melhorar a precisão e eficiência na caracterização de reservatórios. Em particular, os algoritmos de aprendizado de máquina, especialmente os métodos de aprendizado supervisionado aplicados em modelos de regressão, têm sido amplamente utilizados para prever propriedades petrofísicas a partir de dados de perfilagem geofísica. Tradicionalmente, a interpretação desses perfis depende de métodos empíricos e modelos físicos. No entanto, a crescente disponibilidade de dados e o avanço das técnicas de IA oferecem novas oportunidades para aprimorar essa inferência.

4.3.3 Estudo de Caso

O Random Forest pode ser utilizado para prever o tempo de trânsito sônico a partir de outros registros geofísicos, como raios gama, densidade e porosidade neutrônica. Por exemplo, em um estudo de Zhang (2018), o uso do Random Forest para prever propriedades petrofísicas mostrou uma melhora significativa na precisão das previsões em comparação com métodos tradicionais. A abordagem combinou múltiplas árvores de decisão, aproveitando a capacidade do Random Forest de lidar com dados complexos e não lineares.

Outro estudo, conduzido por Al-Handhali em 2023, aplicou o Random Forest em conjunto com técnicas de seleção de atributos para prever registros sônicos sintéticos em formações de gás no norte de Omã. Nesse estudo, o Random Forest, em combinação com eliminação de características menos relevantes, mostrou-se mais confiável e robusto na predição de perfis sônicos quando comparado a outros métodos.

4.4 Benefícios e Desafios

A aplicação de IA na inferência do perfil sônico oferece diversos benefícios significativos. Um dos principais é a precisão melhorada, pois os modelos de aprendizado de máquina são capazes de capturar relações complexas nos dados, algo que é difícil de alcançar com abordagens tradicionais. Além disso, a automação proporcionada pela IA reduz a necessidade de intervenção manual, acelerando o processo de interpretação e análise dos dados. Outro benefício importante é a adaptabilidade dos modelos baseados em IA, que podem ser facilmente atualizados com novos dados, permitindo uma melhoria contínua na precisão das previsões.

No entanto, a aplicação de IA também enfrenta vários desafios. A qualidade dos dados de treinamento é crucial para a precisão dos modelos, e dados ruidosos ou incompletos podem comprometer as previsões. Além disso, alguns algoritmos de aprendizado de máquina requerem recursos computacionais significativos, o que pode ser um obstáculo, especialmente quando aplicados a grandes conjuntos de dados. Outro desafio é a interpretação dos modelos, pois modelos complexos como redes

neurais profundas podem ser difíceis de interpretar, o que pode limitar sua aceitação em ambientes operacionais.

4.5 Aplicação de Modelos Preditivos de Inteligência Artificial na Engenharia de Produção

A aplicação de dados sintéticos em problemas da engenharia de produção tem se mostrado uma abordagem poderosa para lidar com a escassez de dados e aprimorar a eficácia de modelos de *machine learning*. Em contextos onde a coleta de dados reais é limitada ou dispendiosa, os dados sintéticos gerados por técnicas como *Generative Adversarial Networks* (GANs) e modelos baseados em aprendizado profundo permitem criar conjuntos de dados artificiais que mantêm as características estatísticas do original. Isso tem sido essencial para testar e otimizar processos de produção sem comprometer a privacidade ou a segurança dos dados reais (Goyal & Mahmoud, 2024).

Os benefícios dos dados sintéticos na engenharia de produção incluem a detecção de falhas em sistemas de manufatura. Eles permitem a realização de simulações extensivas e o desenvolvimento de soluções para problemas como a previsão de falhas em equipamentos críticos, utilizando dados gerados de forma controlada (SpringerLink, 2022).

CAPÍTULO 5
METODOLOGIA

A revisão bibliográfica foi realizada com base em uma série de publicações, visando o entendimento da evolução geológica e do sistema petrolífero da Bacia de Lagoa Parda Sul, assim como dos dados de perfilagem utilizados e da aplicação de algoritmos de inteligência artificial para inferência das características do subsolo.

5.1 Levantamento do Banco de Dados

Os dados geofísicos e geológicos utilizados são provenientes do acervo de dados públicos terrestres disponibilizados na página virtual Projeto Reate da CPRM (Serviço Geológico do Brasil) em parceria com a ANP (Agência Nacional de Petróleo). A partir de procedimentos de controle de qualidade, verificando-se a quantidade e a qualidade dos dados, foi então escolhido o campo de produção Lagoa Parda Sul para ser estudado.

Das 88 perfurações localizadas no campo Lagoa Parda Sul, que tiveram coleta de dados do perfil dos poços através de sondas de perfilagem, 28 possuem perfil sônico (DT) registrado. Após a exclusão dos poços que não possuem os preditores utilizados 13 puderam ser utilizados para a construção do modelo.

Ao total serão utilizados 15.504 eventos para a construção do modelo, com uma média de 1.192,6 eventos em cada poço. Como é possível observar na tabela 5.1 o poço 4-LP17-ES e 7-LP-8-ES possuem menos de 100 eventos para serem testados, tornando a previsão suscetível a volatilidade em relação aos critérios de avaliação do modelo.

Tabela 1: *Quantidade de eventos disponíveis por poço*

Nome do Poço	Tamanho Amostral
1-LP-54-ES	934
3-LP-60-ES	2232
3-LP-71-ES	1207
4-LP-55-ES	3627
4-LP-86-ES	2212
4-LP-87–ES	1225
4-LP-17-ES	68
7-LP-11-ES	535
7-LP-19-ES	604
7-LP-39-ES	653
7-LP-42-ES	600
7-LP-63-ES	1551
7-LP-8-ES	56

5.2 Definição dos critérios de avaliação do modelo de inferência

Para a avaliação do modelo de inteligência artificial supervisionado de regressão utilizado na inferência do perfil sônico, foram adotados os critérios de coeficiente de determinação (R^2) e raiz do erro quadrático médio (RMSE). A escolha desses critérios fundamenta-se na necessidade de uma análise abrangente e precisa do desempenho do modelo, garantindo que as estimativas das propriedades das formações subterrâneas sejam confiáveis e úteis na prática geofísica.

O coeficiente de determinação (R^2) permite avaliar a capacidade do modelo de capturar a variação observada nos dados do perfil sônico. Um valor de R^2 próximo de 1

indica um bom ajuste do modelo, enquanto um R² nulo indica que o modelo preditivo não consegue explicar nenhuma variabilidade dos dados, ou seja, as previsões do modelo são tão boas quanto simplesmente prever a média dos valores observados. Já um R² negativo significa que o modelo está se ajustando pior do que essa abordagem simples de prever a média.

Coeficiente de Determinação R²:

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} = 1 - \frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2} \qquad \text{Equação (5.2.1)}$$

Componentes das Equações

- y_i: Valor observado
- $\hat{y}_i$: Valor predito pelo modelo
- $\bar{y}$: Média dos valores observados
- n: Número de observações

A raiz do erro quadrático médio (RMSE), por sua vez, foi escolhida por fornecer uma medida absoluta da precisão das previsões do modelo. O RMSE é calculado como a raiz quadrada da média dos quadrados dos erros de previsão, representando a magnitude média dos desvios entre os valores preditos e os valores reais observados. Esta métrica é crucial para avaliar a precisão das estimativas, pois um RMSE baixo indica que as previsões estão próximas dos valores reais medidos, assegurando a confiabilidade das inferências feitas pelo modelo.

Erro Quadrático Médio da Raiz (RMSE):

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2} \qquad \text{Equação (5.2.2)}$$

Componentes das Equações

- y_i: Valor observado
- $\hat{y}_i$: Valor predito pelo modelo
- n: Número de observações

A combinação de R^2 e RMSE oferece uma avaliação complementar do desempenho do modelo. Enquanto o R^2 fornece uma visão geral da capacidade explicativa do modelo, o RMSE indica a precisão das previsões em termos absolutos. A utilização de ambos os critérios é essencial para garantir que o modelo de inteligência artificial não apenas capture bem as tendências gerais dos dados, mas também forneça estimativas precisas e utilizáveis no contexto prático da geofísica.

Este estudo abordará apenas o RMSE e o R^2, porém, há uma clara oportunidade de enriquecer a avaliação da robustez do modelo. A análise de resíduos é uma técnica útil para examinar o desempenho preditivo, verificando a distribuição das diferenças entre os valores reais e os previstos. Isso pode ser feito por meio de um teste de normalidade dos resíduos, que avalia se eles estão distribuídos normalmente com média zero. Em um modelo bem ajustado, os erros de previsão devem seguir essa distribuição, indicando a ausência de fatores que possam introduzir viés no conjunto de dados.

5.3 Padronização, limpeza e tratamento dos perfis de poço.

A preparação dos dados de perfis de poço é uma etapa essencial para garantir a precisão e a confiabilidade de um modelo de inteligência artificial. Este processo envolve a padronização, limpeza e tratamento dos dados, de modo a utilizar apenas informações relevantes e corretas no treinamento do modelo.

5.4 Leitura dos dados

Para a leitura dos arquivos de perfil de poço, foi utilizada a biblioteca *Lasio* em Python, uma ferramenta eficiente e amplamente utilizada para ler arquivos com extensão "Las" (Log ASCII Standard). Esta biblioteca permite extrair informações detalhadas dos arquivos Las de forma estruturada, facilitando o acesso e o tratamento dos dados. A utilização da *Lasio* garante uma leitura precisa e rápida dos dados, permitindo sua posterior análise e tratamento.

5.4.1 Padronização da nomenclatura dos mnemônicos

A padronização dos nomes dos mnemônicos é outra etapa crucial. Diferentes empresas podem utilizar nomenclaturas variadas para os mesmos parâmetros, o que pode causar inconsistências e dificuldades na análise dos dados. Padronizar os nomes dos mnemônicos garante que todos os dados sejam interpretados corretamente pelo modelo. Por exemplo, o mnemônico "DT" deve ser consistentemente utilizado para representar o tempo de trânsito sônico em todos os conjuntos de dados.

5.4.2 Verificação da tipagem dos dados

A verificação do tipo dos dados é um passo fundamental para assegurar a integridade e a consistência das informações. Durante a leitura dos arquivos, foi verificado se os dados estavam no formato correto, garantindo que não houvesse tipos de dados incorretos ou inconsistentes que pudessem comprometer a análise. Além disso, a leitura da descrição dos mnemônicos foi realizada para compreender o significado de cada parâmetro e assegurar que todos os dados estivessem corretamente identificados e padronizados.

5.4.3 Identificação da localização dos poços.

Validar a localização dos poços também é uma etapa crítica no tratamento dos dados. A validação da localização assegura que os dados correspondem ao mesmo

campo, evitando a contaminação do treinamento do modelo em relação a poços provenientes de locais com características geológicas distintas ao local de análise, prevenindo erros na interpretação e análise dos resultados.

5.4.4 Valores faltantes

Primeiramente, foram selecionados apenas os poços que possuem a curva de tempo de trânsito perfil sônico (DT). A presença do perfil sônico é fundamental porque, em um modelo de aprendizado supervisionado, é necessário ter as respostas corretas (ou rótulos) para treinar o modelo. O perfil sônico serve como a variável alvo que o modelo irá aprender a prever a partir dos dados de entrada. Sem esta informação, o modelo não pode ser treinado, impossibilitando a identificação de padrões relacionados ao perfil sônico.

5.4.5 Valores inconsistentes e análise da distribuição

Os histogramas são ferramentas eficazes para a identificação de valores atípicos que podem distorcer a análise dos dados e prejudicar a performance do modelo. A visualização gráfica facilita a detecção desses valores extremos, permitindo que sejam tratados de maneira adequada, seja pela remoção ou transformação, garantindo assim a integridade dos dados utilizados no modelo.

Os histogramas também auxiliam na detecção de valores ausentes ou inválidos. É comum encontrar valores ausentes preenchidos com um valor padrão, como zero ou um número negativo. Estas práticas podem ser facilmente identificadas através de barras altas em torno desses valores específicos nos histogramas. Detectar e tratar esses valores é essencial para assegurar que os dados sejam completos e válidos, evitando possíveis problemas durante o treinamento do modelo.

A verificação da consistência dos dados entre diferentes poços é outra vantagem da utilização de histogramas. Ao comparar os histogramas de mnemônicos de diferentes poços, podemos identificar inconsistências ou erros de medição. Se os mnemônicos de um poço apresentarem uma distribuição muito diferente dos outros,

isso pode indicar a necessidade de revisão e correção dos dados desse poço específico. Garantir a consistência dos dados é fundamental para a robustez do modelo de inteligência artificial.

Na figura 6 é apresentado histogramas dos mnemônicos para o poço 7-LP-63-ES, sendo replicados para todos os poços utilizados na construção do modelo. Este procedimento assegura que os dados sejam visualmente inspecionados quanto à distribuição, outliers, valores ausentes e consistência, permitindo um tratamento adequado. Este passo é fundamental para garantir a qualidade dos dados utilizados no treinamento do modelo, resultando em previsões mais precisas e confiáveis do tempo de trânsito do perfil sônico.

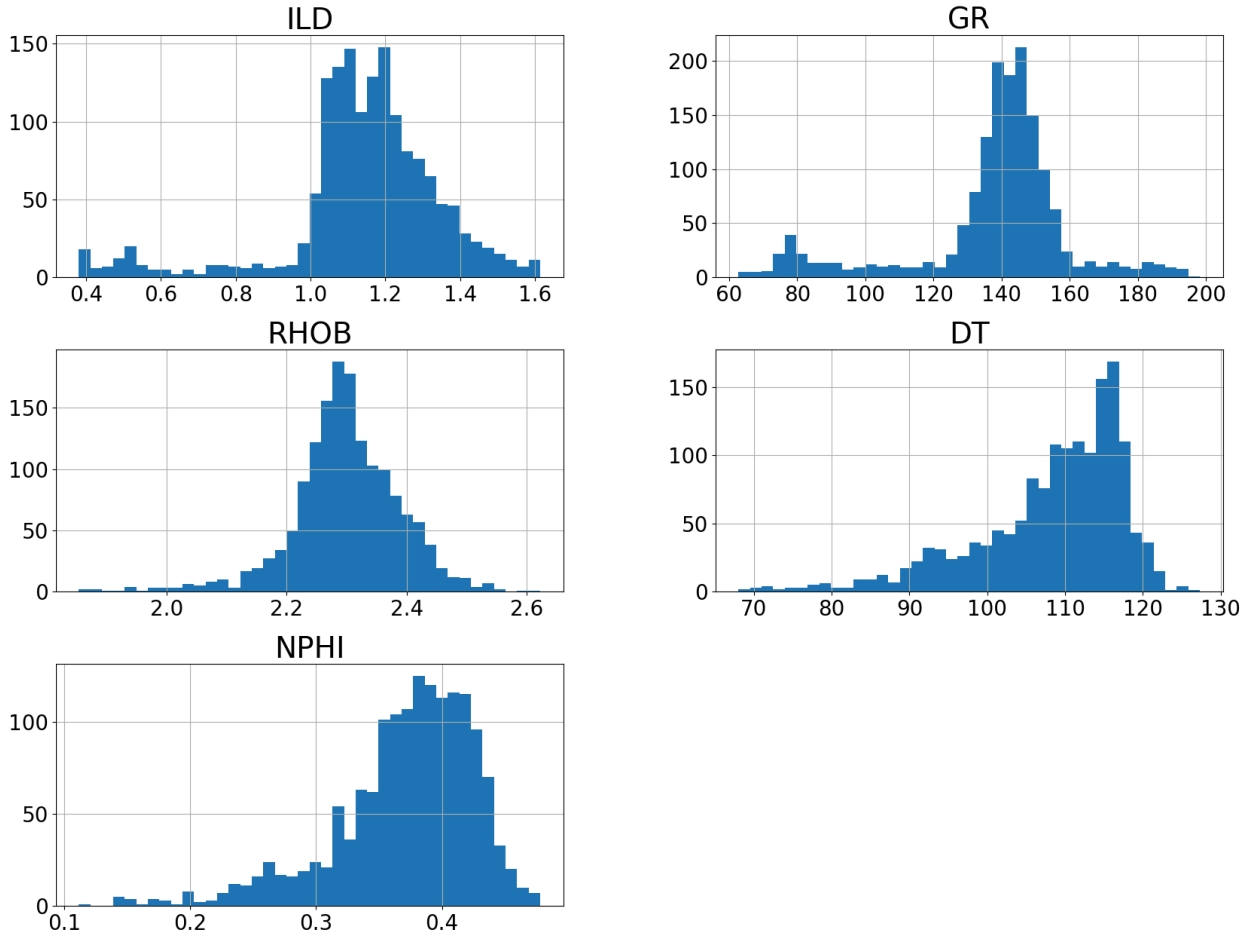

Figura 6: Histograma poço 7-LP-63-ES

A utilização de histogramas na etapa de limpeza e tratamento dos perfis de poço oferece uma visão detalhada dos dados, permitindo a identificação e correção de possíveis problemas. Este processo melhora significativamente a qualidade dos dados e contribui para a construção de um modelo de inteligência artificial mais eficaz e preciso.

Complementando essa abordagem, a utilização de boxplots também desempenharam um papel crucial, especialmente quando se trata de analisar as variáveis que desejamos inferir, como o tempo de trânsito do perfil sônico (DT). Os boxplots fornecem uma visualização detalhada da distribuição dos valores de DT, destacando a mediana, os quartis e os possíveis outliers. Esta análise é fundamental para detectar anomalias e valores atípicos que podem distorcer os resultados do modelo de inteligência artificial.

Assim como demonstrado na figura 7, os boxplots facilitam a verificação da consistência dos dados entre diferentes poços, identificando erros de medição ou variações anômalas que precisam ser corrigidas. A identificação e o tratamento adequado desses outliers e inconsistências antes do treinamento do modelo são cruciais para assegurar a precisão e a robustez das previsões do perfil sônico, contribuindo significativamente para a confiabilidade e a validade do estudo geofísico.

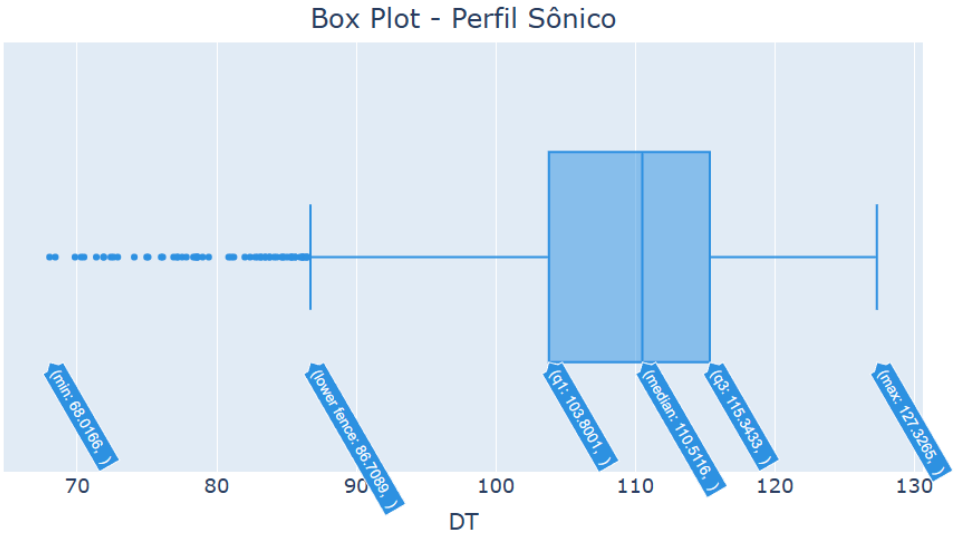

Figura 7: Boxplot poço 7-LP-63-ES

5.5 Identificação dos atributos relevantes para a inferência do tempo de trânsito compressional

A identificação de atributos relevantes para a inferência do tempo de trânsito do perfil sônico (DT) é uma etapa fundamental na construção de um modelo de inteligência artificial robusto e preciso. Uma abordagem eficaz para este fim é a análise de correlação, que permite avaliar a relação entre os diversos atributos dos perfis de poço e o perfil sônico, facilitando a seleção das variáveis mais informativas.

A análise de correlação envolve o cálculo dos coeficientes de correlação de Pearson, que quantificam a força e a direção da relação linear entre duas variáveis. Este coeficiente varia de -1 a 1, onde valores próximos a 1 indicam uma forte correlação positiva, valores próximos a -1 indicam uma forte correlação negativa, e valores próximos a 0 indicam pouca ou nenhuma correlação. Ao aplicar esta técnica aos dados de perfis de poço, é possível identificar quais atributos têm uma correlação significativa com o DT, indicando que esses atributos são potencialmente relevantes para a inferência do DT.

Primeiramente, utilizamos bibliotecas de análise de dados, como *Pandas* e *NumPy* em Python, para calcular os coeficientes de correlação de Pearson entre o perfil sônico e os demais atributos dos perfis de poço. Este passo permitiu quantificar a força das relações lineares entre o perfil sônico e cada atributo, facilitando a identificação dos atributos mais relevantes.

Para facilitar a interpretação dos resultados, criamos uma matriz de correlação que mostra os coeficientes de correlação entre todos os atributos, incluindo o perfil sônico. Esta matriz foi visualizada como um mapa de calor (heatmap) utilizando bibliotecas como *Seaborn* e *Matplotlib*, permitindo uma identificação visual das relações mais fortes. Os atributos que apresentaram coeficientes de correlação mais altos (positivos ou negativos) com o perfil sônico foram selecionados como os mais relevantes. Esses atributos são considerados valiosos, pois suas variações estão fortemente associadas às variações do perfil sônico, sugerindo que contêm informações cruciais para a inferência do perfil sônico.

Além de considerar a força da correlação, também analisamos a direção da correlação. Uma correlação positiva indica que, à medida que o valor do atributo aumenta, o valor do perfil sônico também tende a aumentar, enquanto uma correlação negativa indica o contrário. Compreender essa relação forneceu insights adicionais sobre a natureza dos dados e orientou a modelagem do nosso sistema de inteligência artificial.

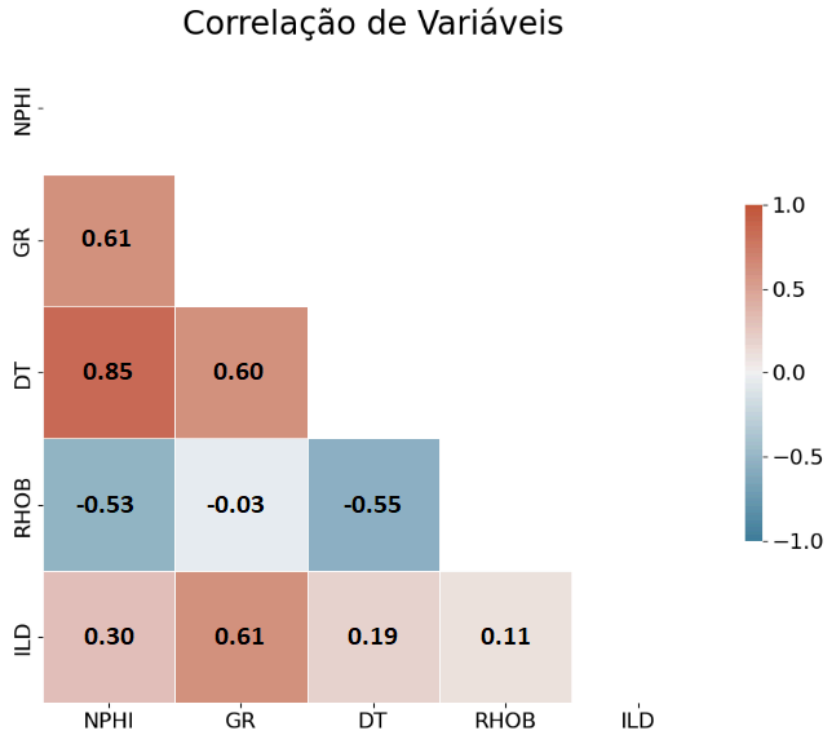

Figura 8: Mapa de calor da correlação dos mnemônicos do poço 7-LP-63-ES

Os resultados da figura 8 demonstram que a análise de correlação destacou os atributos com maior relevância para a inferência do perfil sônico. A eliminação de atributos com baixa correlação ajudou a simplificar o modelo, reduzindo a dimensionalidade e potencialmente melhorando o desempenho e a interpretabilidade do modelo. Assim, a análise de correlação se mostrou uma ferramenta eficaz para identificar os atributos mais relevantes para a inferência do tempo de trânsito do perfil sônico (DT), assegurando que o modelo de inteligência artificial seja treinado com os dados mais informativos e úteis, resultando em previsões mais precisas e confiáveis.

5.6 Criação e Treinamento do Modelo Random Forest para Geração de Perfis Sônicos Sintéticos

Para a inferência do perfil sônico, foi utilizado o algoritmo Random Forest, conhecido por sua robustez e capacidade de lidar com grandes volumes de dados e diversas variáveis. O *Random Forest* é um método que constrói múltiplas árvores de decisão durante o treinamento e retorna a média das previsões individuais das árvores para melhorar a precisão.

5.6.1 Engenharia de Características na Modelagem de Dados

A criação e o treinamento do modelo começaram com a engenharia de características (feature engineering), uma etapa essencial que envolve a criação de novas variáveis a partir dos dados brutos existentes. No contexto deste estudo, criamos novas características, como o identificador único de poço, que permitem ao modelo captar informações específicas e relevantes que não estão diretamente presentes nos dados originais. A engenharia de características ajuda a enriquecer o conjunto de dados, fornecendo mais informações ao modelo, o que pode melhorar significativamente a capacidade de previsão. Ao criar variáveis que capturam melhor a essência dos dados, aumentamos a chance de o modelo identificar padrões e relações importantes que podem ser cruciais para a previsão precisa dos perfis sônicos.

5.6.2 Codificação de Variáveis Categóricas para Modelagem Eficiente

A variável categórica, como nomes de poços, precisa ser transformada em formatos numéricos através de técnicas de codificação, neste caso utilizando a ferramenta *One-Hot Encoding*, para serem processadas pelo algoritmo de machine learning. Este passo garante que o modelo possa interpretar e utilizar todas as informações disponíveis de maneira eficiente. Sem a codificação adequada, o modelo não seria capaz de entender e utilizar a variável categórica do nome do poço, o que poderia resultar em uma perda significativa de informações e em um desempenho inferior do modelo.

5.6.3 Normalização Z-Score na Preparação de Dados para Machine Learning

A normalização (scaling) dos dados utilizando a técnica de escalonamento z-score é uma etapa de extrema importância no pré-processamento de dados para machine learning. O escalonamento z-score transforma os dados ajustando cada variável para que tenha uma média de zero e um desvio padrão de um. Matematicamente, a normalização z-score é expressa pela fórmula:

Z - Score

$$z = \frac{x-\mu}{\sigma} \qquad \text{Equação (5.6.3)}$$

Componentes das Equações:

- z: Z-score
- x: Valor individual da amostra
- μ: Média da população
- σ: Desvio padrão da população

Esta transformação é crucial porque o algoritmo de aprendizado de máquina, *Random Forest*, funciona melhor quando os dados são escalonados de forma uniforme. A normalização evita que variáveis com grandes escalas dominem o modelo, permitindo que todas as características contribuam de maneira equilibrada para o processo de treinamento. Isso resulta em um modelo mais robusto e confiável, com uma maior capacidade de generalização para novos dados. Ao garantir que todas as variáveis estejam na mesma escala, a normalização z-score melhora a precisão e a eficiência do treinamento, além de facilitar a detecção de padrões e relações significativas nos dados.

5.6.4 Separação de Dados em Treinamento, Validação e Teste com Poço Cego

A separação dos dados em conjuntos de treinamento e validação é uma prática essencial em machine learning, fundamental para a construção de modelos robustos e confiáveis. Ao dividir os dados, asseguramos que o modelo seja treinado em um subconjunto dos dados (treinamento) e avaliado em outro subconjunto independente (validação). Esta abordagem permite uma avaliação realista do desempenho do modelo em dados não vistos anteriormente, prevenindo o overfitting, um problema comum onde o modelo se ajusta excessivamente aos dados de treinamento e perde a capacidade de generalização para novos dados.

Além da divisão convencional entre treino e validação, a separação de um poço específico, denominado "blind well", para a validação desempenha um papel crucial no contexto da inferência de perfis sônicos. O poço cego é completamente isolado do processo de treinamento e é reservado exclusivamente para a validação final do modelo. Essa prática permite uma avaliação objetiva da capacidade de generalização do modelo, pois garante que o modelo seja testado em um conjunto de dados que nunca foi utilizado em nenhum estágio do treinamento.

Separar os dados em preditores e a classe alvo também é um passo crítico no pipeline de dados. Os preditores incluem todas as variáveis independentes que serão usadas para prever a variável dependente (classe alvo), que, neste caso, é o perfil sônico. Esta distinção clara entre preditores e a variável alvo garante que o modelo de machine learning seja treinado de forma correta, utilizando todas as informações relevantes disponíveis sem introduzir viés.

5.6.5 Implementação e Otimização do Modelo Random Forest para Perfis Sônicos

Para a construção e treinamento do modelo de inferência de perfis sônicos, foram utilizadas ferramentas da biblioteca *Scikit-Learn*, amplamente reconhecida e utilizada para machine learning em Python. A principal ferramenta empregada foi o algoritmo *Random Forest*, implementado através da classe *RandomForestRegressor*.

O *RandomForestRegressor* foi configurado com 100 estimadores, especificando que o modelo deve construir 100 árvores de decisão. Além disso, definimos uma semente aleatória (random state) para garantir a reprodutibilidade dos resultados.

Utilizar uma semente fixa é crucial para que os experimentos possam ser replicados com os mesmos resultados, assegurando a validade científica das análises.

Para otimizar o modelo e avaliar seu desempenho, utilizamos a técnica de validação cruzada com a ferramenta *GridSearchCV*. Esta técnica realiza uma busca exaustiva sobre os parâmetros especificados, combinada com a validação cruzada, para determinar a melhor configuração dos hiperparâmetros do modelo. O estimador foi definido como o modelo *RandomForestRegressor* configurado anteriormente. A métrica de avaliação utilizada foi o coeficiente de determinação R^2, que mede a proporção da variância explicada pelo modelo. Um valor mais alto de R^2 indica um melhor ajuste do modelo aos dados. Configuramos a validação cruzada com 5 folds, dividindo os dados de treinamento em cinco subconjuntos. O modelo é treinado e avaliado cinco vezes, cada vez utilizando um subconjunto diferente como conjunto de validação e os outros quatro como conjunto de treinamento.

O treinamento do modelo foi realizado utilizando o método fit, que ajusta o modelo aos dados de treinamento. Neste processo, os preditores (variáveis independentes) do conjunto de dados de treinamento e a classe alvo (perfil sônico) foram cuidadosamente preparados. A transformação do vetor de destino em uma forma adequada para o treinamento do modelo foi um passo crucial para garantir a eficácia do processo.

5.7 Valor de referência para avaliação do modelo inferencial

Na modelagem preditiva, o baseline é um ponto de referência que serve para avaliar o desempenho de um modelo. Em termos simples, o baseline pode ser entendido como a performance mínima aceitável ou o padrão inicial contra o qual novos modelos são comparados. Ele é essencial para estabelecer se um modelo proposto oferece melhorias significativas em relação a métodos mais simples ou heurísticas. Segundo Géron (2019), "Um baseline é frequentemente um modelo simples e interpretável, como uma regressão linear, que serve como uma referência para avaliar a eficácia de modelos mais complexos." Na análise de dados, a importância de um baseline reside na sua capacidade de fornecer uma linha de base

objetiva para comparar a precisão, robustez e utilidade dos modelos preditivos desenvolvidos.

A equação de Faust é comumente utilizada como baseline em estudos relacionados à inferência do perfil sônico, especialmente por sua simplicidade e ampla aceitação na indústria. Esse baseline permite que os pesquisadores identifiquem se o uso de técnicas mais avançadas, como o Random Forest, realmente proporciona melhorias significativas. Como ressaltado por James (2013), "Ter um baseline robusto é crucial para evitar a falsa interpretação de melhorias marginais que podem ser, na verdade, resultado de overfitting ou de ruído nos dados". Dessa forma, ao adotar a equação de Faust como baseline, o estudo assegura que o desempenho do modelo seja comparado a um padrão bem estabelecido, garantindo a validade e a relevância dos resultados obtidos.

5.7.1 Definição da Equação de Faust

A equação de Faust é uma fórmula empírica amplamente utilizada na geofísica para estimar a velocidade das ondas sonoras em formações geológicas, a partir de parâmetros de densidade e profundidade. Ela é particularmente útil em ambientes onde os dados diretos de velocidade sônica são limitados ou inexistentes, oferecendo uma estimativa inicial que pode ser refinada com dados adicionais ou métodos mais complexos. Segundo Faust (1953), "a equação foi desenvolvida com base em uma vasta coleção de dados de perfis geofísicos, permitindo estimar a velocidade sônica em função da profundidade e da densidade das rochas." Essa equação tem sido aplicada em diversas áreas da engenharia de petróleo e geologia, servindo como um ponto de partida confiável em muitos estudos de subsuperfície.

A fórmula da equação de Faust é expressa como:

$$V = C \cdot \rho^{m} \cdot D^{n} \qquad \text{Equação (5.4)}$$

Onde:

- V: é a velocidade da onda sônica (em metros por segundo)
- ρ: é a densidade da formação (em g/cm³)
- D: é a profundidade (em metros)
- C, m e n são constantes empíricas derivadas de ajustes com base em dados de campo

Esta equação é valorizada por sua simplicidade e pela capacidade de fornecer resultados razoavelmente precisos em uma variedade de condições geológicas. Como destacado por Schlumberger (1989), "a equação de Faust continua sendo uma ferramenta fundamental em perfis geofísicos, oferecendo uma maneira prática de inferir velocidades sônicas quando medições diretas não estão disponíveis." Embora métodos mais avançados, como o Random Forest, possam superar essa estimativa em termos de precisão, a equação de Faust permanece um *baseline* sólido para validação de modelos mais complexos.

CAPÍTULO 6

RESULTADOS

A Figura 9 contém dois gráficos que comparam os valores reais e preditos do perfil sônico (DT) utilizando o modelo de Random Forest no poço 7-LP-63-ES.

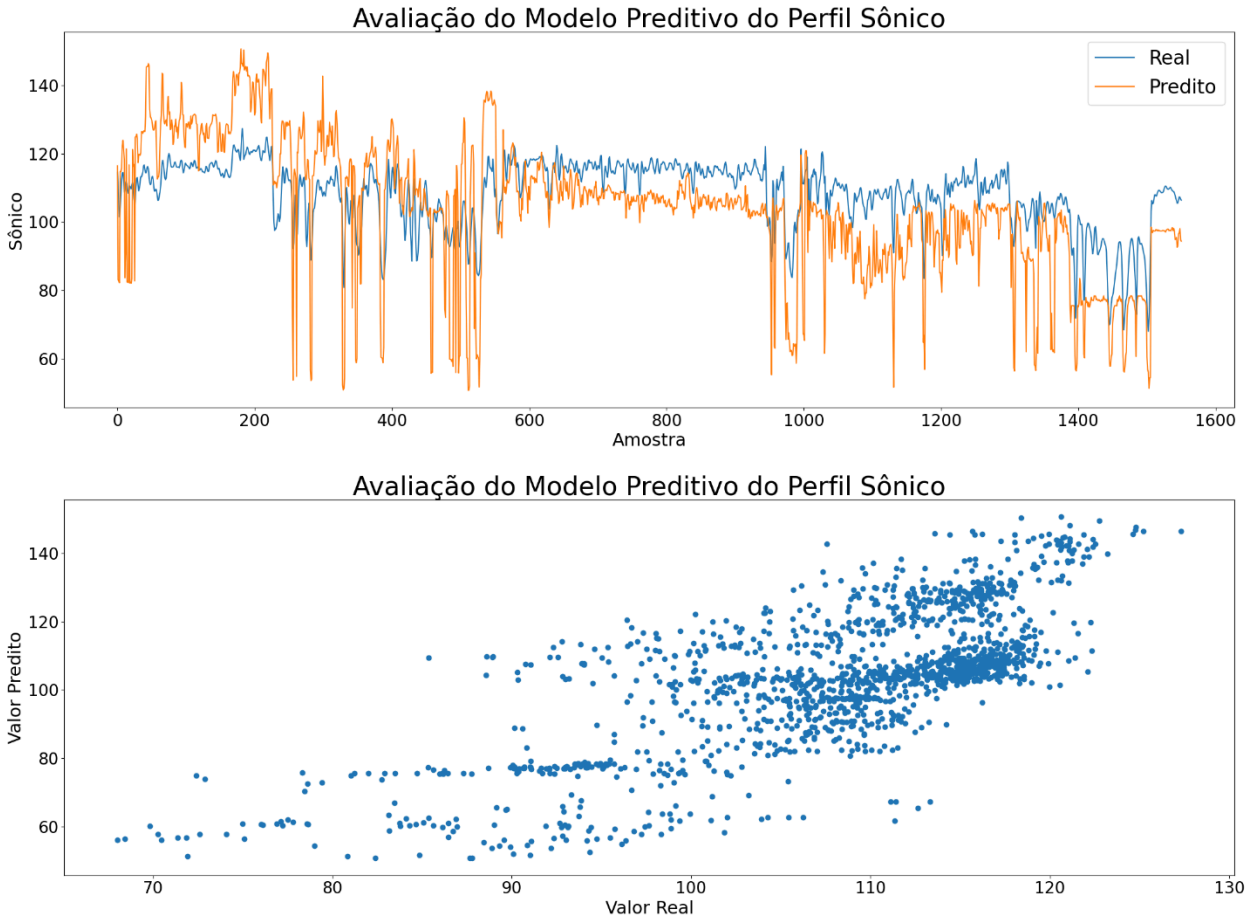

Figura 9: Avaliação do modelo preditivo por gráfico de sequência de amostras e gráfico de dispersão no poço 7-LP-63-ES

6.1 Comparação de Séries Temporais

No gráfico superior, temos a comparação entre os valores reais (curva azul) e os valores preditos (curva laranja) ao longo das amostras. Observa-se que as duas curvas não se alinham perfeitamente, indicando discrepâncias entre os valores preditos pelo modelo e os valores reais. Em algumas regiões, a curva predita está significativamente afastada da curva real, sugerindo que o modelo encontra dificuldades em capturar toda a variabilidade dos dados. Essa inconsistência pode ser um sinal de que o modelo possui potencial para melhorar sua precisão a partir de ajustes nos hiperparâmetros do modelo.

6.2 Dispersão de Valores Reais vs. Preditos

O gráfico inferior da figura 6.1 ilustra a relação entre os valores reais e preditos através de um gráfico de dispersão. Idealmente, em um modelo perfeito, todos os pontos se alinham ao longo da linha $y = x$, indicando previsões exatas. No entanto, o gráfico mostra uma dispersão considerável em torno dessa linha. Especificamente, para valores reais mais altos do perfil sônico, os valores preditos tendem a ser subestimados, enquanto para valores reais mais baixos há uma grande variabilidade. Isso sugere que o modelo apresenta desempenho inconsistente, especialmente nas extremidades dos valores reais do perfil sônico.

6.3 Análise do Desempenho do Modelo Preditivo e Comparação com a Equação de Faust

Na figura 10, observamos a comparação entre o sônico sintético gerado pelo algoritmo *Random Forest* e o valor obtido pela equação de Faust. Avaliar o sônico sintético em relação ao valor coletado pela ferramenta de coleta de dados do poço permite uma análise objetiva do desempenho do modelo inferencial, destacando as melhorias proporcionadas pelo *Random Forest* em comparação ao valor de referência da equação de Faust.

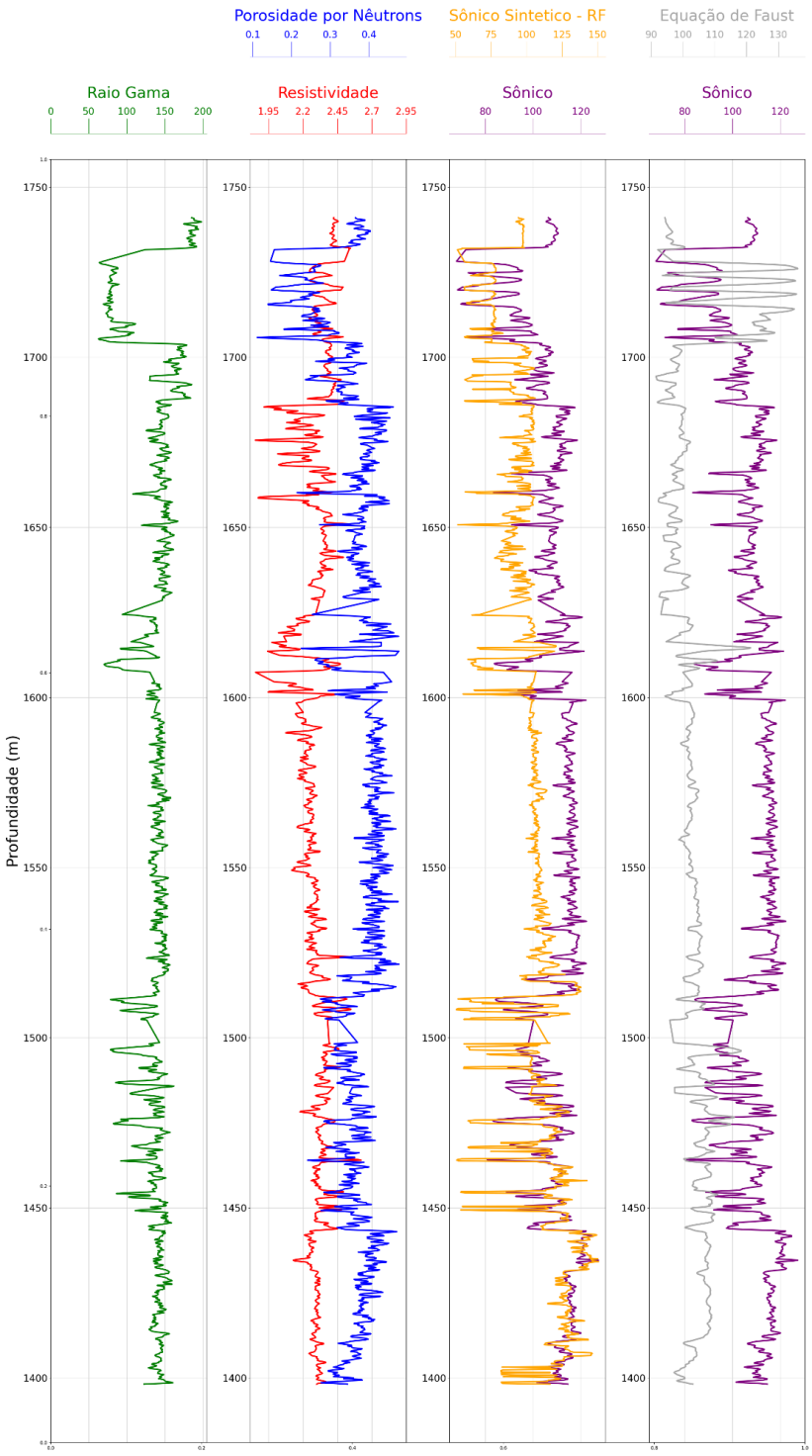

Figura 10: Gráfico de perfis de poço referentes aos mnemônicos Raio Gama, Porosidade por Neutros, Resistividade, Sônico, Sônico sintético pelo modelo de Random Forest e Sônico sintético pela equação de Faust no poço 7-LP-63-ES.

Em várias regiões, os valores preditos pelo modelo *Random Forest* estão mais próximos dos valores reais do que aqueles calculados pela Equação de Faust, sugerindo que o modelo *Random Forest* é capaz de capturar nuances dos dados que a Equação de Faust não consegue.

O desempenho quantitativo do modelo é refletido nas métricas de R^2 e RMSE. O coeficiente de determinação (R^2) de 0.4876 indica que aproximadamente 48.76% da variabilidade dos dados reais é explicada pelo modelo *Random Forest*. Este é um indicador robusto de que o modelo possui um bom ajuste geral aos dados. No entanto, a *Root Mean Square Error* (RMSE) de 15.54 revela que, em média, há uma discrepância considerável entre os valores preditos e os valores reais. Este valor de RMSE sugere que, apesar do modelo *Random Forest* oferecer previsões mais precisas em comparação com a Equação de Faust em várias regiões, ainda existem erros significativos que precisam ser abordados.

Essas discrepâncias podem ser atribuídas a diversos fatores, incluindo a complexidade geológica das formações, a variabilidade intrínseca dos dados de poços e as limitações inerentes ao modelo de *Random Forest* em capturar todas as relações complexas presentes nos dados. Para melhorar o desempenho preditivo, são necessários ajustes adicionais nos hiperparâmetros do modelo, uma engenharia de características mais refinada e a exploração de diferentes técnicas de normalização.

Avaliando o comparativo entre o modelo *Random Forest* e a Equação de Faust, é possível observar que, embora o modelo *Random Forest* demonstra um desempenho superior em várias áreas comparado à Equação de Faust, indicando um avanço na metodologia de inferência de perfis sônicos, há espaço significativo para melhorias. As métricas de desempenho e a análise gráfica sugerem que, com refinamentos adicionais, o modelo preditivo pode oferecer previsões ainda mais precisas e confiáveis, proporcionando um valor agregado substancial na análise geofísica de perfis de poços.

6.4 Interpretação Geral dos Resultados Considerando a Repetição do Experimento para Cada Poço

Os resultados fornecidos na tabela demonstram o desempenho do modelo Random Forest para cada poço individualmente, utilizando métricas de R^2 e RMSE para avaliar a precisão das previsões do perfil sônico. A análise desses resultados revela insights importantes sobre a variabilidade do desempenho do modelo entre diferentes poços.

Tabela 2: Comparação do desempenho do modelo preditivo do algoritmo de regressão Random Forest e equação de Faust através das métricas R^2 e RMSE nos 12 poços avaliados.

Well	R^2		RMSE	
	Faust Equation	Random Forest	Faust Equation	Random Forest
1-LP-54-ES	-0.78	0.46	19.97	24.93
3-LP-60-ES	-0.15	0.62	10.78	27.49
3-LP-71-ES	-0.76	0.10	38.39	30.62
4-LP17-ES	-30.71	0.65	32.66	24.81
4-LP-55-ES	0.18	0.61	17.66	26.75
4-LP-86-ES	0.67	0.62	12.92	19.32
4-LP-87-ES	-0.22	0.43	8.93	36.71
7-LP-11-ES	-15.33	0.25	38.06	16.48
7-LP-19-ES	-4.70	0.32	12.68	20.89
7-LP-39-ES	-2.07	0.11	22.36	19.87
7-LP-42-ES	-1.50	0.14	20.71	16.59
7-LP-63-ES	-1.05	0.48	13.87	15.54
7-LP-8-ES	-12.23	0.31	32.02	23.15

No modelo Random Forest os valores de R^2 variam significativamente entre os poços, com o poço 4-LP-17-ES apresentando o melhor desempenho (R^2 = 0.6536) e o poço 7-LP-39-ES mostrando menor desempenho (R^2 = 0.1124). Isso indica que o modelo Random Forest é capaz de explicar uma proporção moderada da variabilidade dos dados em alguns poços, enquanto em outros, a capacidade preditiva é muito limitada. A métrica RMSE também varia amplamente, com valores que vão de 15.54 a 36.71. Um RMSE menor indica previsões mais precisas, como observado no poço

7-LP-63-ES (15.54), enquanto um RMSE maior, como no poço 4-LP-87-ES (36.71), sugere que as previsões do modelo estão mais distantes dos valores reais, evidenciando a necessidade de melhorias no ajuste do modelo para esses casos específicos.

Ao comparar os resultados do modelo *Random Forest* com a linha de base fornecida pela Equação de Faust, observamos que, em muitos casos, o modelo preditivo oferece previsões mais precisas. Por exemplo, o poço 4-LP-17-ES, com um R^2 de 0.65 e um RMSE de 24.81, demonstra que o modelo Random Forest pode superar a precisão das previsões baseadas na Equação de Faust, que possuem R^2 e RMSE respectivamente -30.71 e 32.66. A variação no desempenho entre os poços pode ser atribuída a diversos fatores, incluindo a complexidade geológica das formações, a qualidade e a quantidade dos dados disponíveis para cada poço, e a presença de heterogeneidades nas formações rochosas. Esses fatores podem influenciar significativamente a capacidade do modelo de capturar as nuances dos dados de perfil sônico.

CAPÍTULO 7
CONCLUSÕES

Os resultados obtidos ao longo deste estudo permitem uma análise detalhada sobre o desempenho do modelo Random Forest para a inferência do perfil sônico em diferentes poços. Enquanto a Equação de Faust, tradicionalmente utilizada como baseline, mostrou-se inadequada em vários casos, especialmente quando observamos os valores negativos de R^2 em poços como 7-LP-11-ES, 7-LP-17-ES e 4-LP-42-ES. Esses valores negativos indicam que a equação não apenas falhou em modelar corretamente os dados, mas também apresentou uma pior performance do que simplesmente utilizar a média dos valores observados. Por outro lado, o *Random Forest*, com exceção de alguns poços, apresentou valores de R^2 significativamente superiores, confirmando sua capacidade de capturar as complexas relações não lineares nos dados.

Ao analisar o erro quadrático médio (RMSE), observamos que, apesar da superioridade do Random Forest em termos de R^2, os resultados não foram uniformemente melhores em todas as situações. Em alguns poços, como o 4-LP-86-ES, o RMSE do Random Forest foi inferior ao da equação de Faust, indicando uma melhor precisão nas previsões. No entanto, em outros poços, como o 3-LP-71-ES, a Equação de Faust apresentou um RMSE menor, o que sugere que, em certas condições geológicas, a simplicidade da equação empírica ainda pode oferecer uma estimativa razoavelmente precisa. Isso reforça a ideia de que, embora o Random Forest seja mais robusto em termos de ajuste do modelo, a equação de Faust pode ser útil em cenários onde a simplicidade e a velocidade do cálculo do perfil sônico sintético são cruciais.

Além disso, é importante destacar que o modelo *Random Forest* não apenas superou a Equação de Faust em termos de R^2 em muitos poços, mas também se mostrou mais consistente em sua performance, com menores variações entre os poços analisados. Essa consistência é particularmente importante em estudos geológicos, onde a variabilidade natural dos dados pode dificultar a aplicação de modelos

preditivos. A capacidade do Random Forest de manter um desempenho robusto através de diferentes cenários geológicos sugere que ele é uma ferramenta valiosa para a inferência do perfil sônico, especialmente quando os dados disponíveis apresentam alta complexidade ou não linearidade.

Vale destacar que existem diversas oportunidades de melhoria que podem ser exploradas para aumentar ainda mais a precisão e a eficácia das previsões. Uma área chave para aprimoramento é a otimização dos hiperparâmetros do modelo *Random Forest*. Ajustes como o número de árvores na floresta, a profundidade máxima das árvores e o número mínimo de amostras para dividir um nó podem influenciar significativamente o desempenho do modelo. Além disso, segmentar os dados de acordo com as formações geológicas pode trazer benefícios substanciais, uma vez que diferentes formações apresentam características específicas que podem ser capturadas com maior precisão quando modeladas separadamente. Outra oportunidade importante é a comparação do desempenho do modelo com outros algoritmos de inteligência artificial, como Redes Neurais, *Gradiente Boosting* ou *Suport Vector Machine* (SVM).

O estudo demonstra que, embora a equação de Faust continue sendo uma referência valiosa, o modelo Random Forest apresenta vantagens significativas, especialmente em cenários mais complexos. No entanto, a escolha do modelo deve ser feita com base em uma análise cuidadosa das condições específicas de cada projeto, equilibrando a necessidade de precisão com as limitações práticas de recursos. Isso garante que a metodologia aplicada seja não apenas tecnicamente robusta, mas também alinhada com os objetivos e restrições de recursos do estudo em questão.

REFERÊNCIAS BIBLIOGRÁFICAS

Almeida, F. F. M. (1977). Geologia do Brasil. São Paulo: Editora Nacional.

Ameen, M. S., Smart, B. G. D., Somerville, J. M., Hammoudah, J., & Collier, R. E. L. (2009). Predicting rock mechanical properties of carbonates using log-derived data: A case study from the Middle East. Marine and Petroleum Geology.

Agência Nacional do Petróleo, Gás Natural e Biocombustíveis. (2021). Sumário Geológico e Setores em Oferta - Bacia do Espírito Santo. Disponível em: ANP.

Assaad, F. A. (2009). Geophysical Well Logging Methods of Oil and Gas Reservoirs.

Asquith, G. B., & Krygowski, D. (2004). Basic well log analysis. AAPG Methods in Exploration Series.

Azeem, T., Yanchun, W., Khalid, P., Xueqing, L., Yuan, F., & Lifang, C. (2016). An application of seismic attributes analysis for mapping of gas bearing sand zones in the sawan gas field, Pakistan. Acta Geodaetica et Geophysica.

Bishop, C. M. (2006). Pattern Recognition and Machine Learning.

Breiman, L. (2001). Random Forests

Domingos, P. (2012). A Few Useful Things to Know About Machine Learning. Communications of the ACM.

Ellis, D. V., & Singer, J. M. (2007). Well Logging for Earth Scientists.

Faust, L. Y. (1953). "A Velocity Function Including Lithologic Variation."

França, R. L., Del Rey, A. C., Tagliari, C. V., Brandão, J. R., & Fontanelli, P. R. (2007). Bacia do Espírito Santo. Boletim de Geociências da Petrobras.

Gamal, H., Alsaihati, A., Elkatatny, S., & Abdulraheem, A. (2021). Sonic Logs Prediction in Real Time by Using Random Forest Technique. ARMA/DGS/SEG International Geomechanics Symposium.

Géron, A. (2019). Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems (2ª ed.).

Goyal, M., & Mahmoud, Q. H. (2024). A Systematic Review of Synthetic Data Generation Techniques Using Generative AI. Electronics.

Gunning, J., & Glinsky, M. E. (2007). Use of linear regression for predicting sonic transit time from gamma, density, and porosity logs.

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

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning with Applications in R. Springer.

Jain, S., & Saraf, D. (2019). Artificial neural networks for predicting porosity from sonic well logs and other data.

Johnson, H. M. (1962). A history of well logging.

Lin, W. (2016). A review on shale reservoirs as an unconventional play–the history, technology revolution, importance to oil and gas industry, and the development future. Acta Geologica Sinica‐English Edition.

McCarthy, J., Minsky, M. L., Rochester, N., & Shannon, C. E. (2006). A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence.

Mitchell, T. M. (1997). Machine Learning. McGraw Hill.

Mohriak, W. U., Nemcok, M., & Enciso, G. (2008). South Atlantic divergent margin evolution: Rift-border uplift and salt tectonics in the basins of SE Brazil. Geological Society, London, Special Publications.

MOHRIAK, W.; SZATMARI, P.; ANJOS, S. M. C. (2008). SAL - Geologia e Tectônica - Exemplos nas Bacias Brasileiras (ORG.).

Rasouli, V., Sutherland, S., & Evans, B. (2011). Application of wellbore stability analysis for planning underbalanced drilling in laminated sandstone formations. Journal of Petroleum Science and Engineering.

Rider, M., & Kennedy, M. (2011). The Geological Interpretation of Well Logs. Rider-French Consulting Ltd.

Russell, S. J., & Norvig, P. (2016). Artificial Intelligence: A Modern Approach.

Saumya, J., Kumar, A., Sain, K., & Raju, S. (2019). Synthetic shear sonic log generation utilizing hybrid machine learning techniques for gas hydrate-bearing sediments. Marine and Petroleum Geology.

Schlumberger. (1972). Log interpretation manual/principles. Schlumberger Well Services, Inc., Ridgefield, Vol. I.

Schlumberger (1989). Log Interpretation Principles/Applications. Schlumberger Educational Services.

Schobbenhaus, C., & Brito Neves, B. B. (2003). Mapa Geológico do Brasil ao Milionésimo. Brasília: CPRM.

SpringerLink (2022). Synthetic data use: exploring use cases to optimize data utility. Discover Artificial Intelligence.

Sutton, R. S., & Barto, A. G. (2018). Reinforcement Learning: An Introduction (2nd ed.).

Valadão, R. C., & Lima, J. F. (2017). Geologia da Bacia do Espírito Santo: uma revisão das características sedimentares e tectônicas. Revista Brasileira de Geociências.

Yu, Yanxiang & Xu (2021). Synthetic Sonic Log Generation With Machine Learning: A Contest Summary From Five Methods.

Zhang, L., Liu, Y., & Li, X. (2018). A machine learning approach to predict lithofacies from well logs in the shale reservoirs of the Sichuan Basin, China. Journal of Petroleum Science and Engineering.

Zhao, T., & Carr, M. H. (2013). Synthetic Sonic Log Generation with Machine Learning: A Case Study.