Universidade do Estado do Rio de Janeiro

Centro de Tecnologia e Ciências

Faculdade de Geologia

Daiane dos Santos Cardoso

Utilização de *Machine Learning* supervisionado para a predição de litofácies no Pré-sal da Bacia de Santos, Formação Barra Velha, Campo de Tupi

Rio de Janeiro

2023

Daiane dos Santos Cardoso

Utilização de *Machine Learning* supervisionado para a predição de litofácies no Pré-sal da Bacia de Santos, Formação Barra Velha, Campo de Tupi

Tese apresentada, como requisito parcial para obtenção do título de Doutora, ao Programa de Pós-Graduação em Geociências, da Universidade do Estado do Rio de Janeiro. Área de concentração: Análise de Bacias.

Orientador: Prof. Dr. Marcus Vinicius Berao Ade

Coorientadores: Prof. Dr. Rodolfo Dino e Dr. Pedro Mário Cruz e Silva

Rio de Janeiro

2023

CATALOGAÇÃO NA FONTE
UERJ / REDE SIRIUS / BIBLIOTECA CTC/C

C268 Cardoso, Daiane dos Santos.
Utilização de *Machine Learning* supervisionado para a predição de litofácies no Pré-sal da Bacia de Santos, Formação Barra Velha, Campo de Tupi. / Daiane dos Santos Cardoso. – 2023.
87 f. : il.

Orientador: Marcus Vinicius Berao Ade.
Coorientadores: Rodolfo Dino e Pedro Mário Cruz e Silva.
Tese (Doutorado) – Universidade do Estado do Rio de Janeiro, Faculdade de Geologia.

1. Geologia estratigráfica – Santos, Bacia de - Teses. 2. Machine learning - Teses. 3. Modelagem geológica - Teses. 4. Prospecção sísmica – Teses. 5. Pré-sal – Teses. I. Ade, Marcus Vinicius Berao. II. Dino, Rodolfo . III. Silva, Pedro Mário Cruz e. IV.Universidade do Estado do Rio de Janeiro. Faculdade de Geologia. V. Título.

CDU: 551.7(815.6)

Bibliotecária Responsável: Priscila Freitas Araujo/ CRB-7: 7322

_____ _____

Assinatura Data

Daiane dos Santos Cardoso

Utilização de *Machine Learning* supervisionado para a predição de litofácies no Pré-sal da Bacia de Santos, Formação Barra Velha, Campo de Tupi

Tese apresentada, como requisito parcial para obtenção do título de Doutora, ao Programa de Pós-Graduação em Geociências, da Universidade do Estado do Rio de Janeiro. Área de concentração: Análise de Bacias.

Aprovada em 28 de Dezembro de 2023.

Banca Examinadora: _____

Prof. Dr. Marcus Vinicius Berao Ade (Orientador)

Faculdade de Geologia - UERJ

_____

Prof. Dr. Rodolfo Dino (Coorientador)

Faculdade de Geologia - UERJ

_____

Dr. Pedro Mário Cruz e Silva (Coorientador)

NVIDIA

_____

Prof. Dr. Sérgio Bergamaschi

Faculdade de Geologia - UERJ

_____

Prof.ª Drª. Karla Tereza Figueiredo Leite

Universidade do Estado do Rio de Janeiro - UERJ

_____

Prof.ª Drª. Karen Maria Leopoldino Oliveira

Universidade Federal do Ceará - UFC

_____

Prof.ª Drª. Rosalia Barili da Cunha

Pontifícia Universidade Católica do Rio Grande do Sul - PUCRS

Rio de Janeiro

2023

DEDICATÓRIA

À minha família e amigas, dedico esta tese. Foi o espírito de colaboração e o apoio constante que vocês me deram que tornaram possível a conclusão desta pesquisa. Cada página desta tese reflete o amor e a força que recebi de vocês.

AGRADECIMENTOS

Agradeço à minha família, cujo amor, apoio e incentivo foram a força motriz por trás desta jornada. Cris, Rejane, Eunice, Nina, Carol, Stephanie e Cristine, vocês foram meu porto seguro e a fonte constante de motivação em cada etapa deste caminho.

Agradeço também às minhas amigas Sabrina, Rosa, Fran, Marina e Patrycia, e aos meus amigos Carlinhos, Adriano e Rodrigo, por sempre oferecerem um sorriso, uma palavra de conforto e momentos de descontração, que foram essenciais para manter meu equilíbrio e sanidade durante os momentos mais desafiadores.

Aos meus professores Marcus e Dino, agradeço imensamente pela orientação e paciência. À professora Karla e ao professor Léo, pela disponibilidade, acolhimento, por compartilhar conhecimento e inspirar a excelência.

Às minhas colegas Tainá e Mauren, agradeço por todo o apoio e compreensão nos dias finais de preparação deste documento. Me faltam palavras para agradecer. Eu não teria conseguido sem vocês. Aos demais colegas do IPR-PUCRS e do projeto Plataforma, Felipe, Clarissa, Priscila, Antônio, Leo, João e Cássia, pelo ambiente colaborativo e pelas inúmeras discussões que enriqueceram tanto minha pesquisa quanto minha experiência profissional.

Aos estudantes e profissionais que compartilharam comigo essa longa jornada, como Pedro Henrique, Marília e Karen, e à professora Heather e aos demais colegas e amigos que fiz em Oklahoma, Karelia, Alex, Diana, Ana, Carol e Letícia, vocês tornaram essa etapa inesquecível.

Agradeço ao LAGESED e ao Projeto PRESAL pela oportunidade e pelo financiamento. Agradeço também ao Programa de Pós-Graduação em Geociências da UERJ e à AAPG *Foundation* (prêmio *Grants-in-Aid*), que contribuíram com parte do financiamento para a realização da minha pesquisa na Universidade de Oklahoma. E um agradecimento especial aos amigos da AAPG que me apoiaram com direcionamentos técnicos Lizbeth e Wendel.

Cada um de vocês desempenhou um papel vital nesta jornada e nas conquistas alcançadas. Minha jornada até aqui não teria sido a mesma sem o apoio e a companhia de todos vocês. Obrigado de coração.

Todas as vitórias ocultam uma abdicação.

Simone de Beauvoir

RESUMO

CARDOSO, Daiane dos Santos. **Utilização de *Machine Learning* supervisionado para a predição de litofácies no Pré-sal da Bacia de Santos, Formação Barra Velha, Campo de Tupi**. 2023. 87 f. Tese (Doutorado em Geociências) – Faculdade de Geologia, Universidade do Estado do Rio de Janeiro, Rio de Janeiro, 2023.

A classificação de litofácies é essencial para a exploração e desenvolvimento de campos de petróleo e gás, fornecendo informações valiosas para análises petrofísicas e sísmicas. No caso das litofácies carbonáticas, a heterogeneidade em diferentes escalas apresenta um desafio, especialmente porque a interpretação manual por especialistas é demorada e sujeita a viés. No Brasil, a descoberta de grandes campos de hidrocarbonetos no intervalo de reservatórios do Pré-sal gerou um alto interesse na origem e classificação de rochas carbonáticas complexas. Recentemente, métodos de machine learning emergiram como uma solução promissora para processos de classificação automatizada de litofácies. Neste estudo, apresentamos uma abordagem que inclui o uso do algoritmo XGBoost e um conjunto de dados da Formação Barra Velha no Campo de Tupi, Bacia de Santos, Brasil, realizando a classificação automatizada de litofácies carbonáticas com alta precisão. A metodologia incluiu seleção de dados, controle de qualidade e pré-processamento, seguidos por aumento de dados, definição de hiperparâmetros para cada modelo, treinamento do algoritmo XGBoost e avaliação do modelo usando métricas importantes nas etapas de treinamento, validação cruzada e teste. Doze modelos diferentes foram avaliados, sendo quatro deles selecionados para a sua aplicação nos blind wells. Esses modelos apresentaram 0,63 a 0,72 na métrica de avaliação de modelos F1-score. Os resultados demonstram a eficácia do algoritmo XGBoost combinado com seleção avançada de variáveis, processo de aumento de dados e definição de hiperparâmetros. A abordagem e o fluxo de trabalho criados com o conjunto de dados da Formação Barra Velha do Campo de Tupi forneceram *insights* valiosos para a exploração do intervalo Pré-sal no Brasil, bem como para a sua promissora utilização na classificação de litofácies em diferentes bacias carbonáticas, como as bacias de Campos e Kwanza.

Palavras-chave: *Machine learning* supervisionado. classificação de litofácies. Fomação Barra Velha. Campo de Tupi. Pré-sal da Bacia de Santos.

ABSTRACT

CARDOSO, Daiane dos Santos. **Utilization of supervised Machine Learning to predict lithofacies in the Pre-salt of the Santos Basin, Barra Velha Formation, Campo de Tupi**. 2023. 87 f. Tese (Doutorado em Geociências) – Faculdade de Geologia, Universidade do Estado do Rio de Janeiro, Rio de Janeiro, 2023.

Lithofacies classification is crucial for the exploration and development of oil and gas fields, providing valuable information for petrophysical and seismic analyses. In the case of carbonate lithofacies, the heterogeneity across different scales poses a challenge, particularly because manual interpretation by experts is time-consuming and prone to bias. In Brazil, the discovery of large hydrocarbon fields in the pre-salt reservoir interval has generated significant interest in the origin and classification of complex carbonate rocks. Recently, machine learning methods have emerged as a promising solution for automated lithofacies classification processes. In this study, we present an approach that includes the use of the XGBoost algorithm and a dataset from the Barra Velha Formation in the Tupi field, Santos Basin, Brazil, to perform high-precision automated classification of carbonate lithofacies. The methodology involved data selection, quality control, and preprocessing, followed by data augmentation, hyperparameter tuning for each model, training of the XGBoost algorithm, and model evaluation using important metrics in the stages of training, cross-validation, and testing. Twelve different models were evaluated, with four of them being selected for their application in blind wells. These models achieved F1-scores ranging from 0.63 to 0.72. The results demonstrate the effectiveness of the XGBoost algorithm combined with advanced variable selection, data augmentation process, and hyperparameter tuning. The approach and workflow created with the dataset from the Barra Velha Formation of the Tupi field provided valuable insights for exploring the pre-salt interval in Brazil, as well as for its promising application in classifying lithofacies in different carbonate basins, such as the Campos and Kwanza basins.

Keywords: supervised *Machine Learning*. lithofacies classification. Barra Velha Formation. Tupi Field. Pre-Salt Santos Basin.

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

UERJ	Universidade do Estado do Rio de Janeiro
IME	Instituto de Matemática e Estatística
ML	*Machine Learning*
ANP	Agência Nacional do Petróleo, Gás Natural e Biocombustíveis
XGBoost	*eXtreme Gradient Boosting*
FBVE	Formação Barra Velha
IBM	International Business Machines Corporation
KNN	*K-Nearest Neighbors*
SVM	*Support Vector Machine*
AdaBoost	*Adaptive Boosting*
LightGBM	*Light Gradient-Boosting Machine*
SOM	*Self-Organizing Map*
UFRJ	Universidade Federal do Rio de Janeiro
ES	Espírito Santo
MG	Minas Gerais
RJ	Rio de Janeiro
SP	São Paulo
PR	Paraná
SC	Santa Catarina
W	Oeste
S	Sul
NE	Nordeste
SW	Sudoeste
NNE	Norte-Nordeste
E	Leste
Ma	Milhões de anos
CAM	Formação Camboriú
PIÇ	Formação Piçarras
ITP	Formação Itapema
ARI	Formação Ariri

FLO	Formação Florianópolis
GUA	Formação Guarujá
BDEP	Banco de Dados de Exploração e Produção
LAGESED	Laboratório de Geologia Sedimentar
PGC	Perfilagem Geofísica Convencional
PGA	Perfilagem Geofísica Avançada
T-CV	Treinamento com Validação Cruzada
GR	Raios Gama
RHOB	Densidade
NPHI	Porosidade Neutrônica
DTC	*Compressional Slowness*
RT-10	Resistividade Rasa
RT-30	Resistividade Média
RT-90	Resistividade Profunda
HFK	GR Espectral - Concentração de Potássio
HURA	GR Espectral - Concentração de Urânio
HTHO	GR Espectral - Concentração de Tório
PEF	Fator Fotoelétrico
AL_WF	ECS - Fração de Peso de Alumínio
CA_WF	ECS - Fração de Peso de Cálcio
IRON_WF	ECS - Fração de Peso de Ferro
SI_WF	ECS - Fração de Peso de Silício
SU_WF	ECS - Fração de Peso de Enxofre
TI_WF	ECS - Fração de Peso de Titânio
TCMR	NMR- Porosidade Total
T2LM	NMR - Média Logarítmica T2
CMFF	NMR - Volume de Fluido Livre
BFV	NMR - Fração Volumétrica de Fluido
CQ	Controle de Qualidade
TVDSS	*True Vertical Depth Subsea*
SMOTE	*Synthetic Minority Over-sampling Technique*
TP	*True Positive*
TN	*True Negative*

FP	*False Positive*
FN	*False Negative*
DP	Desvio Padrão
Q1	Primeiro Quartil
Q2	Segundo Quartil ou Mediana
Q3	Terceiro Quartil
CM	Classe Majoritária
CRE	*Calcarenitic*
SHB	*Shrubstone*
CRS	*Crustone*

LISTA DE SÍMBOLOS

%	Porcentagem
boe	Barril de óleo equivalente
Km^2	Quilômetro quadrado
m	Metro
Km	Quilômetro
=	Igual
+	Soma
÷	Divisão
×	Multiplicação
≥	Maior ou igual
≤	Menor ou igual
cm	Centímetro
km	Quilômetros

SUMÁRIO

INTRODUÇÃO

Novas abordagens continuamente impulsionam a evolução do setor de energia, e a indústria de exploração de hidrocarbonetos é uma das grandes responsáveis pelos avanços tecnológicos nesse setor, além de ser a principal fornecedora do suprimento primário de energia no Brasil, representando cerca de 46% da oferta energética do país (MME; EPE, 2023). Entre essas abordagens, destacam-se as técnicas de *machine learning* (ML), que envolvem a aplicação de algoritmos capazes de desenvolver modelos baseados em dados, facilitando a avaliação e interpretação de enormes quantidades de dados (DHARMIK; BAWANKAR, 2023). A utilização desses bancos de dados é possível devido ao avanço da ciência de dados, que inclui o aumento da capacidade de armazenamento e processamento, e algoritmos mais eficazes e eficientes, que têm como objetivo principal alcançar a funcionalidade ideal com o mínimo de confusão (DHARMIK; BAWANKAR, 2023).

No Brasil, o intervalo de reservatórios situado sob uma extensa camada de sal, denominado Pré-sal, está localizado na margem continental leste e é o responsável pelo suprimento da maior parte da demanda de hidrocarbonetos (ABELHA; PETERSOHN, 2018; ANP, 2023). Esse intervalo ocorre nas bacias do Espírito Santo, Campos e Santos, e, em outubro de 2023, foi responsável por 76,4% da produção nacional total de hidrocarbonetos, totalizando 3,441 milhões de boe/dia. A Bacia de Santos, há algum tempo figura como a mais prolífica do país, contribuindo com 3,350 milhões de boe/dia, o que representa mais de 97% da produção total do Pré-sal em setembro de 2023. O Campo de Tupi, foco deste estudo, lidera a lista dos maiores campos produtores do Pré-sal. Esse campo foi responsável por mais de 32% da produção total do intervalo no mês de referência, correspondendo a 1,086 milhões de boe/dia (ANP, 2023).

Os reservatórios responsáveis pelas importantes reservas do Campo de Tupi são formados predominantemente por litofácies carbonáticas pertencentes à Formação Barra Velha (FBVE), conhecidas por apresentarem grande heterogeneidade devido a seus processos deposicionais e diagenéticos (MOREIRA et al., 2007; BORGHI et al., 2022; DE ROS; OLIVEIRA, 2023; PEREIRA et al., 2023). A determinação precisa de propriedades que controlam a qualidade desses reservatórios, como litofácies, porosidade e permeabilidade, tem um papel significativo na exploração e produção de petróleo. Além disso, a classificação de litofácies é importante para a exploração e desenvolvimento de campos de petróleo e gás,

particularmente por fornecer informações essenciais para análises petrofísicas e sísmicas (ABDOLAHI et al., 2022).

Apesar de sua importância, a caracterização precisa de litofácies carbonáticas é dificultada devido à variabilidade das suas propriedades em diferentes escalas, tornando mais complexa a aquisição de dados e a avaliação petrofísica (LUCIA; KERANS; JENNINGS, 2003; BUST; OLETU; WORTHINGTON, 2011; MALKI et al., 2023). Embora os testemunhos forneçam informações confiáveis, seu custo limita o uso regular. Os dados de calha de perfuração são uma alternativa amplamente utilizada, mas podem fornecer uma classificação pouco acurada das litofácies, limitando sua adequação para a caracterização do intervalo do reservatório. Além disso, a descrição manual de litofácies, um método trabalhoso e demorado (KUMAR; RAO; SEELAM, 2022), pode introduzir tendenciosidades quando aplicada a grandes conjuntos de dados (HALOTEL; DEMYANOV; GARDINER, 2020). As litofácies que ocorrem nos gigantescos reservatórios do Pré-sal apresentam uma série de características únicas, tornando sistemas manuais de classificação complexos e, por vezes, inadequados (DE ROS; OLIVEIRA, 2023). Contudo, apesar da caracterização da incerteza interpretativa na classificação de litofácies ser complexa, sua definição é muito importante para decisões de desenvolvimento de reservatórios (HALOTEL; DEMYANOV; GARDINER, 2020).

Uma solução para mitigar o problema da complexidade e incerteza na classificação de litofácies é a utilização de perfis geofísicos avançados e convencionais, pois eles indicam litofácies de forma qualitativa a quantitativa (RIDER, 2002). Além de indicar litofácies, os perfis geofísicos, quando associados a dados sísmicos, fornecem informações relevantes para a caracterização de reservatórios e oferecem resolução vertical adequada para tal caracterização (ABDOLAHI et al., 2022). A correlação desses dados com informações sísmicas também é uma opção para solucionar o problema da calibração sísmica, fornecendo estimativas precisas da qualidade de reservatórios e identificação de novos intervalos produtivos (RAMADHAN et al., 2019; ABDOLAHI et al., 2022).

Estratégias de ML podem ser aplicadas para superar os desafios na caracterização dos importantes reservatórios do Pré-sal. A aplicação de algoritmos de ML supervisionado auxilia na classificação automatizada de litofácies, oferecendo uma obtenção de dados mais rápida, precisa e com minimização de viés para essa propriedade crucial, desde que os dados sejam de boa qualidade e adequados à tarefa (DHARMIK; BAWANKAR, 2023). Levando em consideração o contexto apresentado, o desafio de pesquisa abordado por esta tese buscou responder: “Como a técnica de ML supervisionada pode aprimorar a classificação de

litofácies predominantemente carbonáticas da FBVE no Campo de Tupi, Bacia de Santos, contribuindo para o aumento da eficiência e precisão da exploração de hidrocarbonetos no Pré-sal brasileiro?".

A tese apresenta os resultados da aplicação do ML supervisionado, que ocorreu através da utilização do algoritmo XGBoost na classificação das litofácies da FBVE no Campo de Tupi, que envolveu a integração de dados de descrição de testemunho e de perfis geofísicos para realizar a classificação automatizada de litofácies da unidade. Essa abordagem incluiu etapas cruciais de pré-processamento de dados, além de estratégias como o aumento artificial de amostras das classes de litofácies, que tiveram um impacto positivo nos resultados dos modelos de classificação. Os modelos de classificação mais bem-sucedidos demostraram desempenho entre 63,17% e 71,87% na métrica de avaliação F1-score.

1 **HIPÓTESES E OBJETIVOS**

Diante da complexidade envolvida na identificação de litofácies carbonáticas do Pré-sal, bem como considerando a limitada quantidade de pesquisas que empregaram técnicas de ML supervisionadas na classificação de intervalos predominantemente carbonáticos, este estudo visou validar duas hipóteses. (1) técnicas de ML supervisionadas podem superar os desafios enfrentados na classificação de litofácies carbonáticas. (2) a aplicação do algoritmo XGBoost pode proporcionar uma classificação mais precisa e com atenuação de viés das litofácies.

A motivação para a realização desta tese está vinculada à importância econômica e energética do intervalo Pré-sal, à disponibilidade de grandes volumes de dados e ao avanço de metodologias que auxiliam no entendimento de características indispensáveis em análises exploratórias. Dessa forma, o objetivo geral deste estudo consistiu em aplicar a técnica de ML e suas metodologias integradas na tentativa de aprimorar modelos geológicos que levem a uma otimização em processos de exploração e produção de hidrocarbonetos, com ênfase no Campo de Tupi, localizado no intervalo Pré-sal da Bacia de Santos.

Assim, este estudo teve como objetivos específicos:

a) Desenvolver fluxos de trabalho eficientes para o tratamento, controle de qualidade e análise de grandes volumes de dados geológicos, visando uma abordagem mais precisa e integrada;

b) Reduzir significativamente o tempo necessário para a análise de dados geológicos complexos;

2 **FUNDAMENTAÇÃO TEÓRICA**

Na presente seção, são abordados os principais temas da tese, o ML supervisionado. Primeiramente, são apresentadas informações gerais a respeito dos conceitos relacionados ao ML, incluindo seu surgimento, desenvolvimento e estabelecimento no final dos anos 90. Além disso, são detalhadas as categorias de ML e os tipos de algoritmos associados a elas, com ênfase no algoritmo XGBoost.

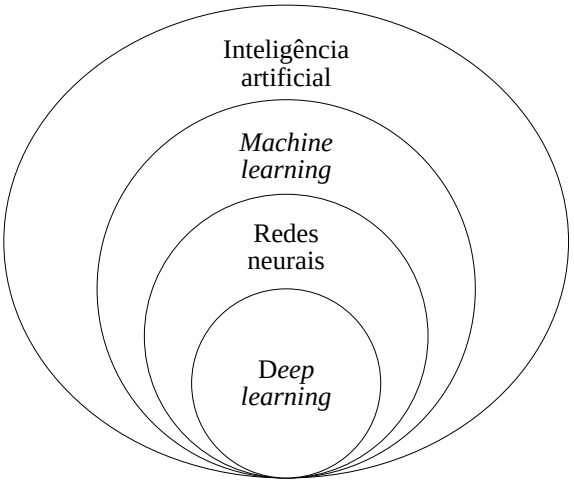
O ML é uma área de pesquisa da Inteligência Artificial que visa o desenvolvimento de programas de computador com a capacidade de aprender a executar uma dada tarefa a partir de própria experiência, utilizando dados históricos (FACELI et al., 2023). O ML abrange as importantes subáreas, como redes neurais e aprendizagem profunda (*Deep learning*) (Figura 1), e consiste em processos de tomada de decisão sequencial sob incerteza, nos quais se utilizam algoritmos especializados para desenvolver programas computacionais de autoaprendizado. Esses programas podem ser utilizados para análise de dados e construção de modelos analíticos. Tais algoritmos podem ser aplicados a uma ampla variedade de problemas, sendo comumente empregados para o reconhecimento de padrões e na descoberta de relações entre conjuntos de dados (BAHL, 2019). Ou seja, consiste no aprendizado sistêmico e automático que se dá através do reconhecimento de características, anomalias e padrões.

O conceito de ML é mais antigo que o surgimento dos primeiros computadores. A ideia de criar máquinas capazes de realizar cálculos de forma autônoma surgiu de matemáticos que viveram no século 17. Um deles foi Blaise Pascal, que criou uma calculadora mecânica capaz de executar operações aritméticas básicas. Outro foi Gottfried Leibniz, que desenvolveu um sistema de números binários que se tornou a fundação do código binário usado em computadores digitais muito mais tarde. Já no século 19, Charles Babbage concebeu os primeiros planos de um dispositivo que poderia ser programado com cartas perfuradas (BAHL, 2019).

O primeiro uso do termo ML é atribuído a Arthur Samuel, um desenvolvedor da International Business Machines Corporation (IBM), nos anos 50, que realizou estudos sobre aprendizado para o jogo de damas (SAMUEL, 1959). Entre os anos de 1960 e 1980, ocorreram outros avanços, como o estudo de reconhecimento de padrões utilizando o método Nearest Neighbor (COVER; HART, 1967), e o desenvolvimento para navegação

independente do carrinho de Stanford, que permitiu sua travessia por uma sala com obstáculos sem intervenção humana (MORAVEC, 1980).

Figura 1 - Relação entre os diferentes métodos do campo de Inteligência artificial.



Fonte: GOMES, 2019. Modificada pela autora, 2023.

Após esse período, ocorreu o chamado “inverno da inteligência artificial”, com pouco desenvolvimento até a retomada no final dos anos 90. Desde então, com avanço das tecnologias e do poder computacional, além do surgimento de bibliotecas de código aberto, sua utilização vem sendo cada vez mais aperfeiçoada, assim como sua capacidade de analisar bancos de dados cada vez maiores. Neste período, podem-se citar como evoluções notáveis a vitória do supercomputador da IBM, Deep Blue, contra o então campeão mundial de xadrez Garry Kasparov em 1997, algoritmos de reconhecimento facial como os desenvolvidos pela empresa Meta em 2014, e aplicações em recentes como carros autônomos (FOOTE, 2021). Atualmente, destacam-se os avanços em *chatbots*, incluindo o conhecido modelo da openAI, ChatGPT.

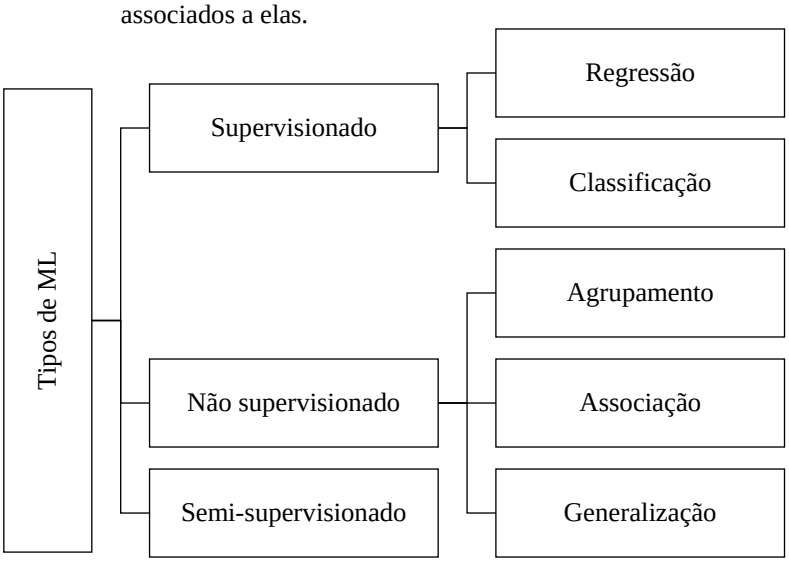
Na jornada de desenvolvimento das técnicas de ML, foram estabelecidas categorias distintas de aprendizado, moldadas pelo nível de interação e supervisão humana que o processo de treinamento necessita. Dessa forma, três categorias de ML foram desenvolvidas: supervisionado, semi-supervisionado e não supervisionado (BAHL, 2019; FACELI et al., 2023) (Figura 2). A categoria supervisionada é a que mais requer interação e supervisão humana, estando associada a algoritmos de classificação e regressão. Para essa categoria, devem ser providos exemplos rotulados e variáveis para o treinamento de algoritmos

específicos (SANCHES, 2003). Além disso, essa categoria exige o acompanhamento contínuo do desempenho desses algoritmos durante o treinamento e teste dos modelos de ML.

Na categoria semi-supervisionada, o nível de interação e supervisão é menor, pois essa categoria consiste na utilização de algoritmos que aprendem a partir de exemplos rotulados e não rotulados (SANCHES, 2003). Nesse caso, os exemplos rotulados são utilizados para a obtenção de informações sobre um dado problema e são utilizados para guiar o processo de aprendizado a partir de exemplos não rotulados (BRUCE, 2001).

Na categoria de ML não supervisionado, o nível de interação é inexistente, ou seja, não há supervisão, estando associada a algoritmos de agrupamento, associação e generalização. Essa categoria é utilizada quando os exemplos não são rotulados e quando utilizam-se algoritmos para prever padrões nos dados de entrada com base em uma característica de similaridade (SANCHES, 2003). A ideia central do ML não supervisionado é agrupar elementos semelhantes sob um mesmo atributo, o que pode revelar grupos que se diferenciam dos demais, facilitando a identificação de características ocultas (ALPAYDIN, 2014).

Figura 2 - Categorias de machine learning e tipos de algoritmos comuns associados a elas.



Fonte: GOMES, 2019. Modificada pela autora, 2023.

A abordagem de ML supervisionado utilizada no desenvolvimento desta tese envolveu um método de classificação, ou modelagem preditiva de classificação, que tem como objetivo o uso de algoritmos para classificar dados com base em suas características. Para que a

máquina seja capaz de atribuir classes aos dados, é necessário que um usuário forneça ao algoritmo pares de dados de entrada e saída conhecidos, normalmente apresentados na forma de vetores. Estes dados podem representar um valor numérico ou uma classe, podendo envolver duas classes (classificação binária) ou diversas classes (classificação multiclasse) (FONTANA, 2020). A modelagem preditiva de classificação aproxima uma função de mapeamento (f) das variáveis de entrada (X) para variáveis de saída discretas (Y). Para realizar esse processo, são empregados conjuntos de dados previamente categorizados, que facilitam o mapeamento de funções às saídas desejadas (FONTANA, 2020).

A generalização, capacidade do algoritmo de atingir bons resultados em dados não utilizados em treinamento, é uma qualidade bastante almejada em modelos supervisionados. Devido a isso, uma das técnicas empregadas é a validação destes modelos treinados, ou aferição do erro, aplicando-o em um conjunto de teste independente, não usado durante o treinamento do modelo (MURPHY, 2012). Com foco nos conjuntos de dados de treinamento e validação, tem-se que a divisão desses conjuntos entre proporções como 80% para a fase de treinamento e 20% para validação é chamada de validação pelo método Holdout (MURPHY, 2012). Outro método de validação mais robusto é o k-fold cross validation, onde um conjunto de dados é dividido em k subconjuntos (folds) e, então, o processo de treinamento e validação acontece k vezes, sempre com um subconjunto diferente sendo utilizado para validação, sendo que a média das métricas de erro para cada k-fold é considerada como a estimativa de erro do modelo (MURPHY, 2012). Na etapa seguinte, os conjuntos de dados de treinamento e validação devem passar pelo processo de definição de hiperparâmetros, treinamento, regularização e validação do treinamento. O processo de definição de hiperparâmetros tem como objetivo ajustar as características do algoritmo adotado para que o modelo treinado não acarrete duas situações indesejadas, *underfitting* e *overfitting*. No *underfitting*, a precisão pode ter baixos resultados e o modelo pode necessitar de reforma para melhorar as previsões. Uma solução possível para isso é aumentar o tempo de treinamento (CARDOSO, 2020). Na situação de *overfitting* o modelo acerta com precisão as predições, porém, quando aplicado ao conjunto de dados teste, tem baixa precisão, tendo em mão um modelo que basicamente decorou o treino. Uma possível solução para esse problema é uso do recurso chamado regularização, ou ainda a adoção de estratégias de validação (CARDOSO, 2020).

Entre os métodos clássicos de algoritmos classificação, destacam-se as árvores de decisão, *k-nearest neighbors* (KNN), *Support Vector Machine* (SVM) e redes neurais (FONTANA, 2020; GOÉS, 2023). A evolução dos algoritmos de ML tem sido constante, expandindo sua aplicabilidade em diversas áreas. Um exemplo dessa evolução é o *Ensemble*

Learning, que emprega múltiplos modelos menos robustos, *weak learners*, na resolução de um mesmo problema. Esses modelos são combinados para melhorar a acurácia, generalidade e robustez em relação a um único *weak learner* (GOÉS, 2023).

Os métodos de *ensemble learning* podem ser divididos em duas categorias, homogêneos e heterogêneos. Nos homogêneos, apenas um tipo de modelo de ML é utilizado para construir o *ensemble*, enquanto nos heterogêneos, diversos modelos de ML são empregados. Entre as metodologias aplicadas na categoria homogênea, destacam-se o *Bagging* e o *Boosting* (GOÉS, 2023). O *Boosting* é particularmente relevante quando se considera a evolução dos algoritmos clássicos de classificação. Este tipo de algoritmo é chamado de *boosting* devido à sua capacidade de impulsionar o poder preditivo de algoritmos menores utilizados na análise do conjunto de dados (BAHL, 2019). Ele se caracteriza pelo treinamento sequencial de seus *weak learners*, focando nas instâncias mais desafiadoras do conjunto de treinamento. A abordagem específica do processo de *Boosting* varia conforme o algoritmo escolhido, sendo o AdaBoost (*Adaptive Boosting*) e o *Gradient Boosting* dois dos principais (GOÉS, 2023).

O algoritmo XGBoost (*eXtreme Gradient Boosting*), uma versão derivada e otimizada do Gradient Boosting, combina centenas de árvores de decisão simples para formar um modelo mais preciso e eficiente. Além disso, proporciona uma abordagem de impulsionamento paralelo, resolvendo problemas de ciência de dados de maneira rápida e acurada (ZHANG; ZHAN, 2017). Este algoritmo emprega combinações de algoritmos mais simples, como os clássicos de ML de classificação, para fazer predições.

O método XGBoost é reconhecido por sua alta velocidade e desempenho, além da capacidade de aproveitar o processamento dos computadores multicore modernos. Esse método tem habilidade de lidar com o treinamento de grandes conjuntos de dados e suporta problemas de classificação multiclasse (FRIEDMAN, 2001; CHEN; GUESTRIN, 2016; KHANDELWAL, 2020). Além disso, usa gradiente descendente para gerar novas árvores a partir das anteriores, direcionando a função objetivo para o seu mínimo. O método combina múltiplas árvores simples com baixa precisão para criar um modelo mais preciso (ZHANG; ZHAN, 2017).

A aplicação do algoritmo XGBoost em problemas de classificação de litofácies é uma abordagem relativamente nova. Entre os primeiros trabalhos que utilizaram este algoritmo para este fim, destacam-se os de Zhang e Zhan (2017) e Merembayev, Yunussov e Yedilkhan (2018). Os primeiros utilizaram o XGBoost em um problema de classificação de litofácies siliciclásticas e carbonáticas provenientes de um reservatório de gás localizado nos Estados

Unidos. Os autores empregaram como variáveis de entrada para o modelo dados de perfis geofísicos convencionais e de contexto geológico, apresentando como resultado um F1-score igual a 0,56. Já no outro estudo, os autores experimentaram utilizar este algoritmo e outros, como *Logistic Regression*, *Random Forest* e K-*Nearest Neighbor*, para a classificação da estratigrafia e litologia de rochas de aquíferos arenosos localizados no Cazaquistão. Eles utilizaram dados de perfis geofísicos convencionais e de elevação, e apontaram que o XGBoost seria o algoritmo mais recomendado para este tipo de problema proposto.

A partir de então, diversos outros estudos aplicaram XGBoost com o objetivo de utilizar dados de perfis para determinação de litologias ou litofácies (DEV; EDEN, 2019; AL-MUDHAFA, 2020; SUN et al., 2020; GU et al., 2021; MEREMBAYEV et al., 2021; HE; GU; XUE, 2022; GAVIDIA et al., 2023). Estes estudos apresentaram resultados variados, conforme as classes estabelecidas para as áreas de estudo, o tipo de dados utilizados e as técnicas de pré-processamento e otimização de parâmetros do algoritmo. Além do algoritmo XGBoost, os métodos supervisionados *Decision Trees*, *Random Forest*, KNN e *LightGBM* também foram explorados em alguns estudos como estratégia para comparação de resultados e seleção do melhor algoritmo para essa aplicação, sendo que a maioria dos trabalhos analisados indica que os modelos com melhores resultados foram aqueles que utilizaram XGBoost.

Dentro do escopo da aplicação do XGBoost para classificação de litologias ou litofácies de reservatórios carbonáticos, destaca-se o estudo de Gavidia et al. (2023), pois a área de estudo analisada foi também o intervalo Pré-sal brasileiro na Bacia de Santos e FBVE. Neste estudo, os autores buscaram classificar nove litofáceis distintas (*anhydrite*, *calcimudstone*, *dolomudstone*, *silicimudstone*, *spherulistone*, *shrubstone*, *grainstone/rudstone*, *volcanic grainstone* e *volcanic rudstone*). Algumas das classes consideradas representam, na verdade, um agrupamento de subclasses encontradas no local. Além do algoritmo XGBoost, o trabalho também apresenta a aplicação dos métodos Random Forest e Self-Organizing Map (SOM) ao conjunto de dados. Embora o SOM seja normalmente utilizado em tarefas de aprendizagem não-supervisionada, no estudo foi utilizado com o objetivo de incorporar os dados de entrada para calibrar um mapa de variáveis e estabelecer correlações entre os dados de entrada (GAVIDIA et al., 2023).

Como atributos de entrada para os modelos gerados, além de perfis geofísicos usualmente utilizados em problemas desse tipo, os autores incluíram variáveis relacionadas aos volumes mineralógicos, como argila, quartzo, calcita e dolomita. Essas variáveis são provenientes de uma etapa anterior de caracterização e modelagem petrofísica avançada,

baseada nos dados dos perfis geofísicos. Neste trabalho, as métricas de desempenho, acurácia e F1-score, ultrapassaram 90% para o algoritmo XGBoost. No entanto, é importante destacar a complexidade envolvida no pré-processamento dos dados de entrada para o modelo gerado, o que exige um tempo de processamento elevado, podendo ser um fator limitante na classificação automatizada de litofácies encontradas nessa área de estudo. Sendo assim, percebe-se que a necessidade de desenvolver uma metodologia automatizada, rápida e eficiente para a classificação de litofácies de reservatórios carbonáticos do Pré-sal, na Bacia de Santos, permanece como um campo de pesquisa em aberto.

3 **CONTEXTO GEOLÓGICO DA BACIA DE SANTOS**

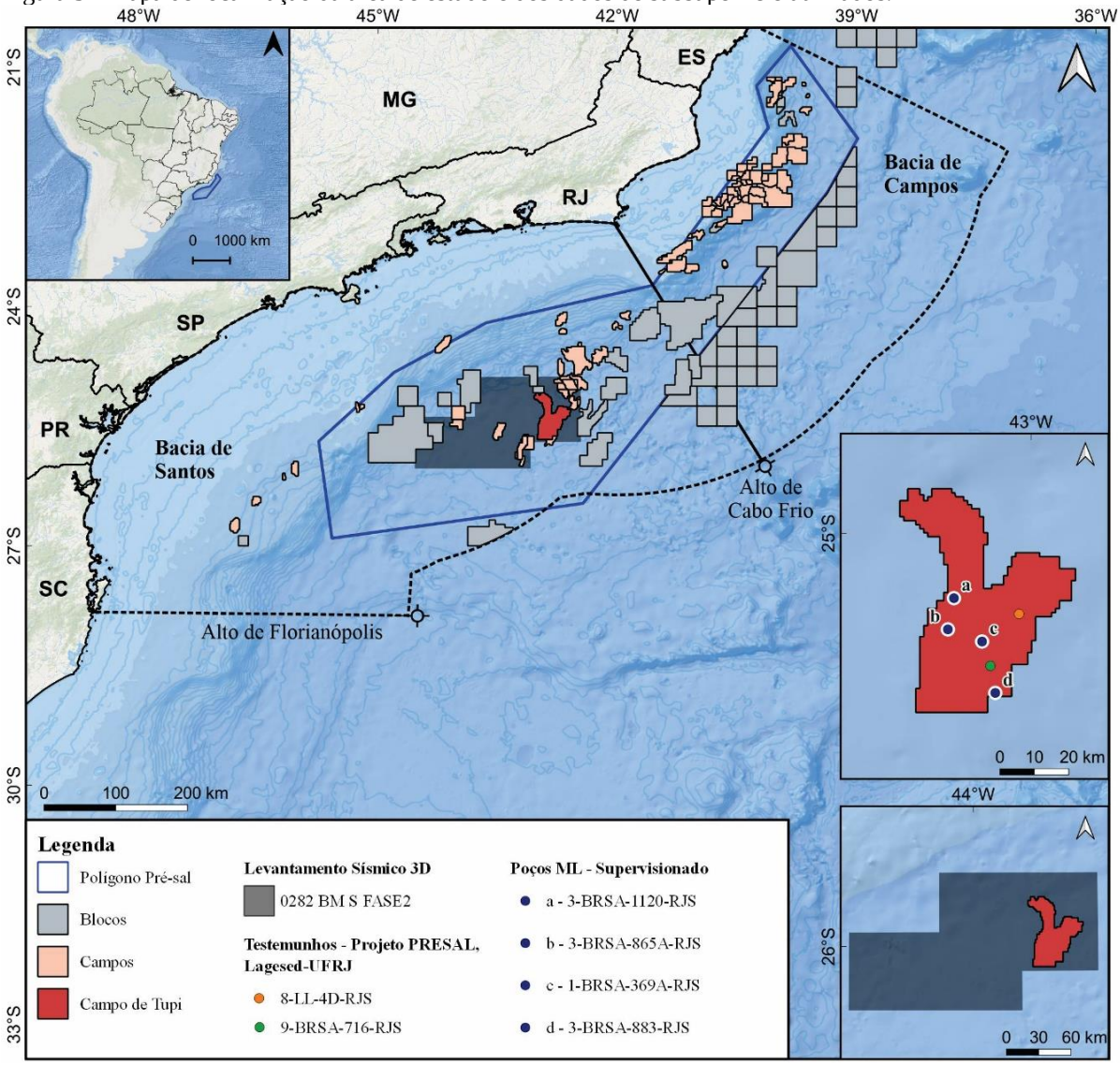
A Bacia de Santos está inserida na porção sudeste da margem continental brasileira, abrange cerca de 350 mil km² e cota batimétrica de até 3.000 m. Ela estende-se pelos litorais dos Estados do Rio de Janeiro, São Paulo, Paraná e Santa Catarina, sendo limitada ao norte pelo Alto de Cabo Frio e ao sul pelo Alto de Florianópolis (Figura 3). A gênese da Bacia de Santos está associada com os processos de rifteamento durante a abertura do Atlântico Sul, no Mesozoico. A sedimentação ocorreu inicialmente em ambientes flúvio-lacustres e progrediu para o estágio de bacia evaporítica até evoluir para uma bacia de margem continental (CHANG et al., 2008).

3.1 **Síntese do contexto geotectônico**

A Bacia de Santos, assim como as bacias de Campos e Espírito Santo, corresponde a uma bacia de margem passiva, formada durante a ruptura do supercontinente Gondwana e abertura do Atlântico Sul no Cretáceo (ALMEIDA et al., 2013). Esta abertura ocorreu de forma diácrona, ao longo de diversos estágios evolutivos, durante os períodos Cretáceo e Jurássico na porção meridional, estendendo-se para o Jurássico e Triássico na porção setentrional (STOLLHOFEN et al., 1998; COBBOLD, PETER ROBERT; MEISLING, KRISTIAN E; MOUNT, 2001; MOHRIAK et al., 2002).

A atuação da tectônica extensional culminou no estiramento da litosfera, no desenvolvimento de falhas na crosta e em processos de rifteamento, levando à geração de bacias sedimentares do tipo rifte na margem sudeste brasileira (MILANI; THOMAZ FILHO, 2000; MOHRIAK, 2012a, 2012b). Sob este regime tectônico, é comum a atividade magmática intensa, bem como a formação de um centro de espalhamento oceânico, com extrusão de magmas basálticos, promovida pela movimentação divergente das placas (MOHRIAK, 2012a).

Figura 3 - Mapa de localização da área de estudo e dos dados de subsuperfície utilizados.



Legenda: unidades federativas do Brasil: ES- Espírito Santo; MG- Minas Gerais; PR- Paraná; RJ- Rio de Janeiro; SC- Santa Catarina; SP- São Paulo.
Fonte: A autora, 2023.

Atualmente, acredita-se que o processo de rifteamento tenha ocorrido sobre estruturas pré-existentes do embasamento, com a abertura do Atlântico Sul seguindo uma direção aproximadamente NE (MOHRIAK et al., 2002; MOULIN; ASLANIAN; UNTERNEHR, 2010; MOHRIAK, 2012a). A formação deste rifte teve início de sul para norte, controlado pela distribuição de tensões regionais, com formação de falhas de direção NNE (CHANG et al., 2008). Devido à extensão oblíqua em relação às estruturas pré-existentes, predominantemente NE-SW, desenvolveram-se falhas de transferência com orientação NW-SE (RIGOTI, 2015).

Quatro modelos geotectônicos são propostos para a formação de bacias sedimentares do tipo rifte, considerando os mecanismos de deformação extensional e a tectônica divergente. Destacam-se os modelos de cisalhamento puro (MCKENZIE, 1978), cisalhamento simples (WERNICKE, 1985), delaminação (LISTER; ETHERIDGE; SYMONDS, 1986), além do de exumação mantélica (LAVIER; MANATSCHAL, 2006). Enquanto os três primeiros são considerados riftes passivos, com magmatismo resultante do estiramento crustal, o último é visto como um rifte ativo.

Independentemente do modelo tectônico, o desenvolvimento de sistemas rifte é caracterizado pela movimentação de falhas normais lístricas, que provocam o abatimento e rotação dos blocos superiores, inclinando-os até acomodar o deslocamento da falha principal (VAN DER PLUIJM, BEN; MARSHAK, 2004). Esta geometria gera sistemas de falhas normais subparalelas que, em seções transversais, se assemelham a peças de dominó caídas (*domino faults*) (LISTER; DAVIS, 1989). Além disso, os processos extensionais na crosta podem levar ao desenvolvimento de uma variedade complexa de estruturas, como falhas escalonadas, *rollover folds*, zonas de imbricação, duplexes extensionais e falhas extensionais de baixo ângulo (VAN DER PLUIJM, BEN; MARSHAK, 2004).

A presença de heterogeneidades no embasamento é uma das causas para a disposição estrutural do sistema rifte da bacia, com seus principais horsts e grábens sendo controlados por falhas normais (ZALÁN et al., 2011). Os lineamentos de direção NE-SW são os mais expressivos, tanto no embasamento quanto nas estruturas presentes no sistema rifte, indicando um processo de rifteamento polifásico (ZALÁN et al., 2011). Além disso, o rifteamento, que ocorreu de forma oblíqua em relação à deriva continental, foi responsável pela formação de uma bacia extensa (MOULIN; ASLANIAN; UNTERNEHR, 2010). Essa atividade tectônica é responsável pela atual geometria da bacia e, possivelmente, pela reativação de estruturas durante a fase sag do intervalo, originando a arquitetura identificada em importantes reservatórios localizados nessa fase.

Além da formação de bacias do tipo rifte, a sua sedimentação e evolução também são controladas pela arquitetura estrutural estabelecida a partir das movimentações tectônicas atuantes no período em questão. Segundo Moreira et al. (2007), a evolução tectonoestratigráfica da Bacia de Santos pode ser dividida em três supersequências principais: Rifte, Pós-Rifte (sag) e Drifte. A fase rifte compreende os sedimentos depositados durante o processo de ruptura do Gondwana e rifteamento. A supersequência pós-rifte engloba a FBVE e a Formação Ariri, depositadas em ambiente transicional entre continental e marinho raso

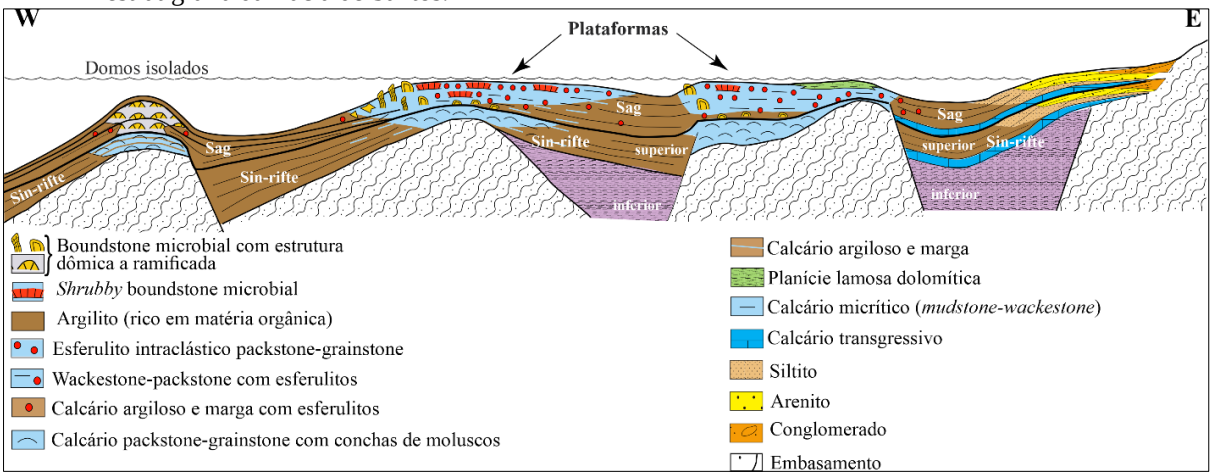
estressante. A supersequência denominada de fase drifte tem sedimentação de origem marinha relacionada à subsidência termal.

As bacias sag, formadas durante o estágio pós-rifte, são geralmente divididas em interior e marginal. Nas bacias sag intracratônicas, a subsidência ocorre por contração termal da litosfera devido a alterações na fonte de calor, enquanto nas bacias rifte a subsidência mecânica é controlada pelas tensões trativas e geração de falhas extensionais. Nichols (2009) define as bacias sag intracratônicas como depressões relacionadas à ampla subsidência da crosta, caracterizadas por serem extensas em área e pouco profundas. A evolução das bacias sag pode ser simples, contendo apenas um ciclo de deposição, ou complexa, com sucessivos ciclos deposicionais.

Na Bacia de Santos, a sequência pós-rifte, também conhecida como fase *Sag*, é representada pela FBVE Superior, enquanto a parte inferior da formação foi gerada durante a fase rifte (WRIGHT; BARNETT, 2015). A fase sag da bacia contém sucessões carbonáticas não marinhas, com evidências de ambiente lacustre. A seção carbonática do tipo rifte-sag, representada pela FBVE, é conhecida como sequência do Pré-sal devido à sua posição estratigráfica (WRITHT; BARNETT, 2017).

Diversos modelos análogos são propostos para os segmentos das margens divergentes no Atlântico Sul, entre as bacias do leste brasileiro e oeste africano, devido à grande semelhança das sequências tectonossedimentares. Essas analogias aplicam-se também aos sistemas petrolíferos e reservatórios distribuídos nas diversas sequências estratigráficas (MOHRIAK, 2012a, 2012b). De forma similar à camada de Pré-sal da FBVE na Bacia de Santos, a seção do Pré-sal da Bacia de Kwanza contém camadas depositadas durante o estágio sin-rifte, que é sucedido pela sedimentação do estágio sag e, por fim, é capeada por camadas de sal (SALLER et al., 2016) (Figura 4).

Figura 4 - Modelo estratigráfico esquemático do Pré-sal na Bacia de Kwanza, Angola: seção análoga à estratigrafia da Bacia de Santos.



Fonte: SALLER et al., 2016. Modificado pela autora, 2023.

3.2 **Estratigrafia**

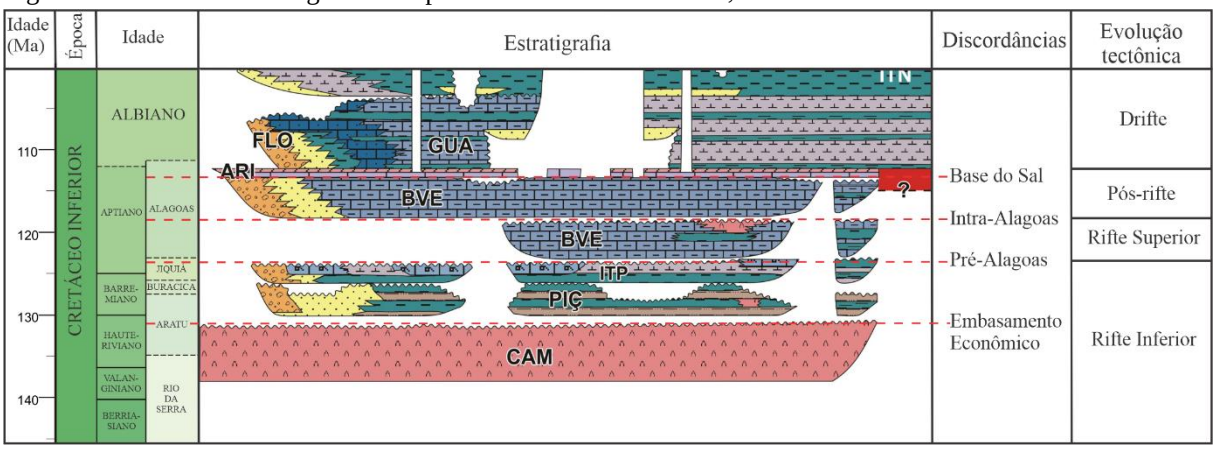
O intervalo estudado ocorre nas Supersequências Pós-rifte e Rifte, inseridas no Grupo Guaratiba e representadas por três sequências deposicionais: K44, K46-48 e K50. As reservas de óleo de interesse para produção ocorrem na FBVE, formada pelas sequências deposicionais K44 e K46-48, depositadas do Eoaptiano ao Neoaptiano (MOREIRA et al., 2007). A FBVE pode ser subdividida em um intervalo inferior (Rifte Superior) e um intervalo superior (Pós-rifte), delimitados por três discordâncias: Pré-Alagoas, limite inferior; Intra-Alagoas, interna; e a localizada no topo, base do sal (MOREIRA et al., 2007; WRIGHT; BARNETT, 2015) (Figura 5).

A Sequência deposicional K44, correspondente à porção inferior da FBVE, foi depositada durante o Eoaptiano e é equivalente ao andar local Alagoas Inferior. Seu limite inferior é marcado pela discordância Pré-Alagoas, onde se observa uma mudança litológica expressiva, passando de depósitos bioclásticos da Formação Itapema para depósitos químicos, originalmente interpretados como “microbialitos” (MOREIRA et al., 2007). Este intervalo é caracterizado litologicamente por calcários microbiais, estromatólitos e laminitos nas porções proximais e por folhelhos nas porções distais. Também ocorrem grainstones e packstones compostos por fragmentos de estromatólitos e bioclastos de ostracodes associados. Tais carbonatos são, por vezes, parcial ou totalmente dolomitizados (MOREIRA et al., 2007).

O limite entre as sequências K44 e K46-48 é marcado pela discordância sísmica interna à formação, chamada de Intra-Alagoas (WRIGHT; BARNETT, 2015; NEVES et al.,

2019; GOMES et al., 2020; BARNETT et al., 2021). A Sequência deposicional K46-48, superior da FBVE, foi depositada durante o Neoaptiano e é equivalente ao andar local Alagoas Superior. Seu limite superior é marcado pela discordância correspondente à base dos evaporitos da Formação Ariri, que marca a ingressão inicial de águas marinhas e atua como um selo para os reservatórios da seção Pré-sal. Essa sequência superior da FBVE é caracterizada, litologicamente, por calcários microbiais intercalados a folhelhos, incluindo calcários estromatolíticos e laminitos microbiais, localmente dolomitizados. As porções proximais à FBVE são compostas por leques aluviais de arenitos e conglomerados (MOREIRA et al., 2007).

Figura 5 - Carta cronoestratigráfica simplificada da Bacia de Santos, detalhando o intervalo de interesse.



Legenda: Formações geológicas: CAM- Camboriú; PIÇ- Piçarras; ITP- Itapema; BVE- Barra Velha; ARI- Ariri; FLO- Florianópolis; GUA- Guaruja.
Fonte: MOREIRA et al., 2007; WRIGHT; BARNETT, 2015. Modificados pela autora, 2023.

Os carbonatos da FBVE são extremamente heterogêneos, variando em origem e textura. Eles podem ser divididos em rochas formadas originalmente por processos químicos (*in situ*) e rochas formadas por processos redeposicionais (retrabalhadas). Estes dois tipos ocorrem com alta frequência de intercalação, inclusive em escala milimétrica, e possuem uma distribuição heterogênea ao longo da bacia. A ausência de análogos com as mesmas condições de formação do lago sag no registro geológico dificulta as interpretações sobre a origem e a gênese desses depósitos. Consequentemente, devido às suas características únicas e à alta complexidade, as litologias que compõem a formação foram descritas e categorizadas de diferentes maneiras por diversos pesquisadores (MOREIRA et al., 2007; TERRA et al., 2010;

FARIAS et al., 2019; GOMES et al., 2020; SILVA et al., 2021; BORGHI et al., 2022; DE ROS; OLIVEIRA, 2023).

Na classificação proposta por Gomes et al. (2020), foram considerados os três componentes singenéticos básicos presentes nesses depósitos: matriz lamosa (argilas magnesianas), esferulitos de calcita e shrubs fasciculares. Já Borghi et al. (2022) consideraram algumas características do modelo de Gomes et al. (2020), sugerindo uma nova classificação. Os autores propuseram o termo "crustone" para rochas semelhantes a crosta, onde os shrubs ocorrem com alto grau de coalescência. Eles também consideraram classes em que há ocorrência simultânea de processos *in situ* e de redeposição, como, por exemplo, o "*calcarenitic crustone*" (crustone calcarenítico). Rochas com quantidades semelhantes de agregados esferulíticos e fasciculares de calcita foram denominadas "*tupilites*" (tupilitos).

Recentemente, De Ros e Oliveira (2023) revisaram as classificações propostas na literatura e propuseram uma nova classificação por meio de diagramas triangulares. No diagrama das rochas *in situ*, as classes correspondem à proporção original entre os três componentes dos vértices: matriz de tamanho argila (essencialmente composta por estevensíta ou outros argilominerais silicáticos), esferulitos de calcita e shrubs de calcita. Os nomes das classes propostas podem ser acompanhados de características como estrutura primária (estratificada, laminada, maciça, etc.), fábrica (coalescida ou não-coalescida) e principais alterações e modificações (parcialmente dolomitizada, silicificada, recristalizada, etc.).

3.3 **Campo de Tupi**

O Campo de Tupi está localizado na porção central do polígono do Pré-sal na Bacia de Santos, há 300 km da costa (TERRA; FERREIRA; OLIVEIRA, 2014) (Figura 3). O campo foi descoberto em 2006 através da perfuração do poço homônimo, em uma campanha *offshore* que teve como objetivo a identificação de alvos em águas profundas na margem continental leste do Brasil. Essa descoberta representou um marco significativo na exploração de petróleo e gás na região, pois revelou a presença de uma das maiores reservas de petróleo já encontradas no Brasil, localizada na Bacia de Santos, na área do Pré-sal (ABELHA; PETERSOHN, 2018).

A produção comercial do campo de Tupi começou em 2010, com a entrada em operação da plataforma P-34, localizada a cerca de 240 km da costa do estado do Rio de

Janeiro. Os primeiros prospectos comerciais com quantidade significativa de hidrocarbonetos foram perfurados no Campo de Tupi, que confirmou a magnitude do sistema petrolífero do intervalo (ABELHA; PETERSOHN, 2018). Ao longo dos anos, a produção aumentou gradualmente, consolidando a posição do Brasil como um importante produtor de petróleo e gás no cenário internacional (ABELHA; PETERSOHN, 2018; ANP, 2023; MME; EPE, 2023)

No ano de 2014, o esforço de perfuração no intervalo Pré-sal levou à descoberta de 23 campos, com Tupi já ocupando o primeiro lugar (MATIAS et al., 2015). Atualmente, esse campo aparece em primeiro lugar na lista dos campos com maior produção acumulada, com 3.367 milhões de barris de óleo equivalente no mês de outubro de 2023. Essa produção acumulada é proveniente de 60 poços, dos quais oito estão entre os maiores produtores de petróleo e gás do intervalo Pré-sal no mês de referência (ANP, 2023).

4 **MATERIAIS E MÉTODOS**

A presente tese foi desenvolvida através da utilização de um conjunto de dados constituído por poços exploratórios e de desenvolvimento, além de descrições de dois testemunhos. Os dados dos poços foram fornecidos pelo Banco de Dados de Exploração e Produção da Agência Nacional do Petróleo, Gás Natural e Biocombustíveis (BDEP-ANP), e as descrições dos testemunhos pelo projeto PRESAL, Laboratório de Geologia Sedimentar (LAGESED-UFRJ), desenvolvido em parceria com a Shell Brasil.

No seu desenvolvimento, foi em empregada uma abordagem de ML supervisionada no problema de classificação de litofácies. Para solucionar o problema, dados de litofácies provenientes da descrição dos dois testemunhos foram utilizados em conjunto com perfis geofísicos como dados de treinamento, desenvolvimento dos modelos de classificação automatizada de litofácies, técnica de ML supervisionada.

4.1 **Base de dados utilizada**

Os cinco poços exploratórios e um poço de desenvolvimento selecionados para a aplicação do ML supervisionado apresentaram perfis geofísicos convencionais e avançados, que indicam parâmetros físicos (resistivos, acústicos e radioativos) e químicos. Dois desses poços, 8-LL-4D e 9-BRSA-716-RJS, além dos perfis geofísicos, também incluem dados de descrições de testemunhos realizadas por pesquisadores do projeto PRESAL, os quais foram fornecidos para o desenvolvimento deste estudo (Tabela 1).

Tabela 1 - Conjunto de dados de poços do Campo de Tupi.

Poço	Tipo de poço	Testemunho	PGC	PGA	Função
1-BRSA-369A-RJS	Exploratório	x	✓	✓	*Blind well*
3-BRSA-865A-RJS	Exploratório	x	✓	✓	*Blind well*
3-BRSA-883-RJS	Exploratório	x	✓	✓	*Blind well*
3-BRSA-1120-RJS	Exploratório	x	✓	✓	*Blind well*
8-LL-4D-RJS	Desenvolvimento	✓	✓	✓	T-CV e teste
9-BRSA-716-RJS	Exploratório	✓	✓	✓	T-CV e teste

Legenda: PGC- perfilagem geofísica convencional; PGA- perfilagem geofísica avançada; T-CV: treinamento com validação cruzada; informação presente (✓); informação ausente (x).

Fonte: A autora, 2023.

4.1.1 Organização das descrições de testemunhos

Os dados de litofácies utilizados no desenvolvimento do estudo compreendem as descrições de testemunhos do poço exploratório 9-BRSA-716-RJS e do poço de desenvolvimento 8-LL-4D-RJS. Esses dados foram revisados, compilados, organizados para se adequarem a resolução dos perfis geofísicos, com um dado de litofácies a cada 0,1524 m. Ambos os poços perfuraram a FBVE no Campo de Tupi, com o poço 9-BRSA-716-RJS representando uma espessura de 170,99 m da formação e o poço 8-LL-4D-RJS representando uma espessura de 99,97 m da formação, correspondendo a 1.778 amostras com informações sobre suas litofácies, as quais foram utilizadas como dados de entrada para os modelos.

Este estudo utilizou as descrições baseadas no sistema de classificação proposto por Borghi et al. (2022). Estas descrições apresentaram 21 litofácies distintas, predominantemente carbonáticas. Do conjunto de amostras, 1.761 são representadas por litofácies como *breccia*, *crustone*, *crystalline limestone*, *dolostone*, *calcarenite*, *calcilutite*, *calcirudite*, *shrubstone* e *spherulstone*. Além dessas litofácies carbonáticas, um conjunto de dados, composto por 17 amostras, incluem *claystone* e *chert* (Tabela 2). As litofácies descritas ocorrem, por vezes, interdigitadas e, por isso, são apresentadas como tipos distintos, dependendo de suas composições principais e secundárias. Neste estudo, para manter a classificação original proposta pelos autores, os nomes compostos das classes de litofácies são apresentados de forma que o primeiro termo representa a litofácies que ocorre de maneira subordinada, enquanto o segundo termo indica a litofácies predominante (Tabela 2).

Tabela 2 - Litofácies caracterizadas a partir da descrição dos testemunhos.

ID	Litofácies	Amostras
1	*Breccia*	10
2	*Chert*	10
3	*Claystone*	7
4	*Crustone*	18
5	*Calcarenitic crustone*	10
6	*Spherulitic crustone*	4
7	*Calcilutite crustone*	5
8	*Crystalline limestone*	4
9	*Dolostone*	3
10	*Calcarenite*	533
11	*Calcirudite calcarenite*	62
12	*Calcilutite calcarenite*	19
13	*Spherulitic calcarenite*	12
14	*Calcilutite*	295
15	*Spherulitic calcilutite*	3
16	*Calcirudite*	96
17	*Crustose*	6
18	*Shrubstone*	338
19	*Calcarenitic shrubstone*	20
20	*Spherulitic shrubstone*	42
21	*Spherulstone*	281

Fonte: BORGHI et al., 2022. Adaptado pela autora, 2023.

4.1.2 Perfis geofísicos

Para a realização desta tese, foram selecionados 21 perfis geofísicos de caráter convencional a avançado, com a seleção baseada na capacidade de indicar litofácies de forma qualitativa, semiquantitativa ou quantitativa, a partir de características físico-químicas das rochas (RIDER, 2002; PEQUENO, 2019).

Os perfis convencionais selecionados incluíram raios gama, sônico, resistividade, densidade e porosidade neutrônica. Os perfis avançados selecionados foram o fator

fotoelétrico, raios gama espectral, espectroscopia de raios gama induzida e ressonância magnética (Tabela 3).

Quadro 1 - Principais usos dos perfis geofísicos.

Usos	Geologia Geral						Geologia de Reservatório		Geoquímica		Petrofísica					Sísmica	
	Litologia	Vulcânicas } litologias não-convencionais	Evaporitos }	Identificação mineral	Correlação estratigráfica	Fácies deposicional	Identificação de fraturas	Identificação sobre pressão	Identificação da rocha geradora	Maturidade	Porosidade	Permeabilidade	Argilosidade	Água salgada de formação	Saturação em hidrocarboneto	Identificação de gás	Velocidade do intervalo	Impedância acústica
Resistividade	-		-		-	-		+	+	+	+	-		-	*	-		
Raios gama	-	-	+	-	-	-			+				+					
Raios gama espectral	-	-	+	+	-	-	-		+				+					
Sônico	+		-		-		+	+	+		*					-	*	* *
Densidade	+	-	-	-		-	+	-	+		*					-		*
Fator fotoelétrico	+	-	-	+														
Porosidade neutrônica	+	-	-	-		-			-		*		-			-		
Ressonância magnética					-	-												
Perfis de imagem	-			-	-	-	+					+			+			
Espectroscopia RG ind.	*	+	+	*					+									

Legenda: Uso essencialmente qualitativo (-); uso semiquantitativo a quantitativo (+); uso estritamente quantitativo (*).

Fonte: RIDER, 2002. Adaptado pela autora, 2023.

Os perfis geofísicos são dados obtidos através do registro de parâmetros físicos e/ou químicos das formações geológicas. Esses registros contínuos de parâmetros ao longo do poço são plotados em função da profundidade medida, originando os perfis geofísicos. Os parâmetros físicos (resistividade, acústica, radioatividade) e químicos medidos podem ser correlacionados com dados geológicos e geofísicos. A correlação entre esses dados permite a produção de resultados precisos e abrangentes, sendo essenciais para a construção de modelos geológicos (CLENEAY, 1992; DOYEN, 2007).

Tabela 3 - Perfis geofísicos selecionados: mnemônicos e utilizações em modelos de ML.

Perfis geofísicos	Mnemônicos
Raios gama	GR
Densidade	RHOB
Porosidade neutrônica	NPHI
Compressional slowness	DTC
Resistividade rasa	RT-10
Resistividade média	RT-30
Resistividade profunda	RT-90
GR espectral - concentração de potássio	HFK
GR espectral - concentração de urânio	HURA
GR espectral - concentração de tório	HTHO
Fator fotoelétrico	PEF
ECS - fração de peso de alumínio	AL_WF
ECS - fração de peso de cálcio	CA_WF
ECS - fração de peso de ferro	IRON_WF
ECS - fração de peso de silício	SI_WF
ECS - fração de peso de enxofre	SU_WF
ECS - fração de peso de titânio	TI_WF
NMR - porosidade total	TCMR
NMR - Média logarítmica T2	T2LM
NMR - volume de fluido livre	CMFF
NMR - fração volumétrica de fluido	BFV

Legenda: GR- raios gama; ECS- raios gama espectral; NMR- ressonância magnética nuclear.
Fonte: A autora, 2023.

4.1.3 Compilação de dados

A compilação de dados envolveu um conjunto de abordagens sistemáticas e organizadas que abrangem a aquisição, organização e análise de informações relevantes para a condução da pesquisa. Essa metodologia desempenha um papel fundamental na obtenção de resultados confiáveis e válidos, permitindo tomar decisões informadas e obter *insights* relevantes (FONTELLES et al., 2010). Dessa forma, após o recebimento dos dados, foram realizadas etapas preliminares para a organização, estruturação e preparação dos dados para análise. Essas etapas preliminares abrangeram a limpeza dos dados, eliminação de erros,

padronização de formatos e a criação de uma estrutura que permitisse a realização de análises, transformando conjuntos de dados desorganizados em informações ordenadas e prontas para interpretação.

Para identificar e caracterizar dados essenciais para este estudo, foi realizada a compilação de dados para a realização da etapa de controle de qualidade. Estes dados incluíram a localização da boca do poço, a profundidade da lâmina d'água, a espessura da mesa rotativa, a orientação e o desvio do poço, entre outras informações contidas na pasta do poço, no perfil composto e no arquivo geral do poço.

4.1.4 Controle de qualidade dos dados

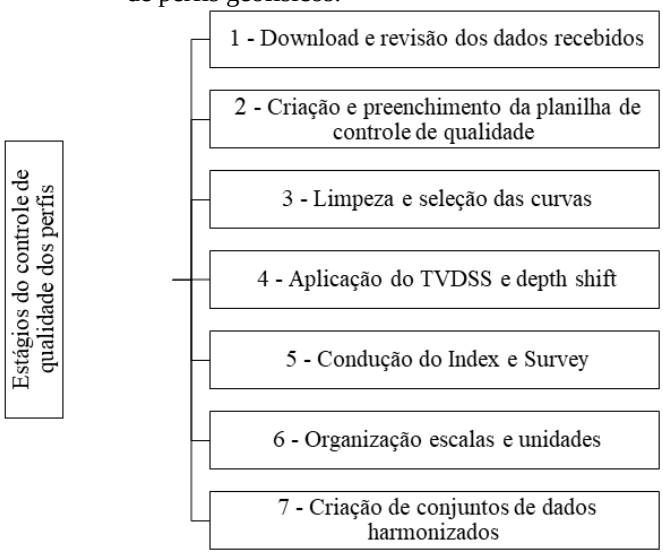
O controle de qualidade (CQ) é um processo essencial que envolve a avaliação e correção de dados para garantir sua confiabilidade e precisão. Esse processo tem como objetivo mitigar erros derivados dos instrumentos, do operador e dos processos de medida envolvidos nas aquisições de dados em subsuperfície (FALAVIGNA et al., 2014). Os dados de perfis geofísicos passaram por uma avaliação para identificar inconsistências ou anomalias, que apontou a necessidade de aplicação do CQ. Esse processo, executado desde o recebimento dos dados até a geração de conjuntos de dados utilizados como dados de entrada, contou com a organização e conferência dos dados e criação de uma planilha de CQ. Essa planilha apresenta informações sobre a perfuração do poço, profundidade dos topos de formação, tipos de dados disponibilizados na pasta do poço e perfilagens corridas (Figura 6).

As etapas seguintes envolveram o carregamento dos arquivos de perfis digitais no software Techlog para a verificação, limpeza e seleção dos perfis geofísicos. Após a seleção dos perfis, foram conduzidas correções relacionadas a suas profundidades. As principais funções relacionadas às correções de profundidade são o *true vertical depth subsea* (TVDSS), *depth shift* e *index* e *survey* (Figura 6). O TVDSS é uma medida de profundidade verdadeira que leva em consideração a inclinação e a direção do poço. As medidas são calculadas a partir da profundidade da mesa rotativa e lâmina d'água, e de tabelas de dados direcionais contidos na pasta dos poços, fornecendo uma medida mais precisa da profundidade em relação a um ponto de referência vertical. O *depth shift* é utilizado para a correlação do posicionamento dos picos de perfis presentes em diferentes *datasets*, para garantir que os perfis geofísicos estejam corretamente alinhados com a profundidade verdadeira. O perfil de raios gama do dataset do

perfil de resistividade foi utilizado como referência para a correção da profundidade dos picos dos outros perfis. Após a correlação dos picos, uma planilha de correção foi gerada e aplicada aos demais *datasets*. As funções *index* e *survey* referem-se ao deslocamento de um perfil geofísico em relação a um índice de referência.

A etapa seguinte consistiu na padronização de escalas e unidades de cada perfil. Por fim, foram gerados *datasets* harmonizados, assegurando que todos os perfis de um mesmo poço estivessem contidos em um único conjunto de dados (Figura 6). A partir desse ponto, os dados de perfis geofísicos de cada poço tornaram-se aptos para uso como dados de entrada em modelos de ML, bem como em análises sísmicas e geofísicas.

Figura 6 - Fluxograma do controle de qualidade dos dados de perfis geofísicos.



Fonte: A autora, 2023.

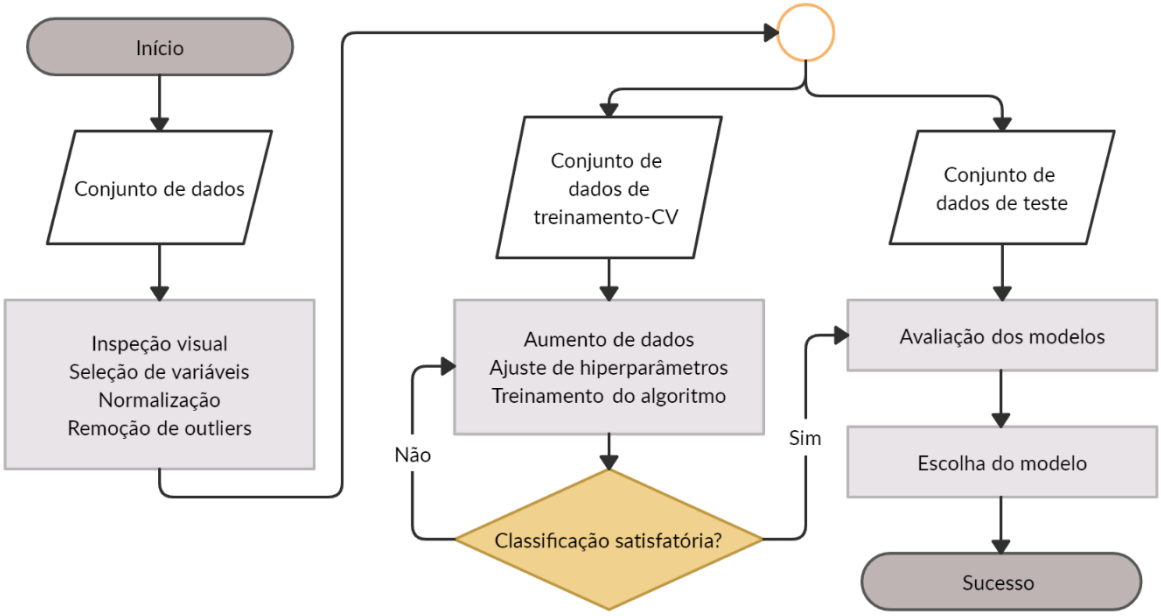
4.2 **Machine learning**

Nas seções subsequentes, são detalhados os aspectos metodológicos relacionados à abordagem de *machine learning* utilizada nesse estudo. Essa abordagem envolveu a aplicação de um algoritmo de aprendizagem supervisionada em uma tarefa de classificação de litofácies predominantemente carbonáticas da FBVE, seção Pré-sal da Bacia de Campos.

Dessa forma, a aplicação da metodologia do presente estudo envolveu diversas etapas, conforme ilustrado na Figura 7. Estas etapas incluíram o pré-processamento e processamento

dos dados, ajuste de hiperparâmetros, treinamento, validação e teste dos modelos, assim como a avaliação e seleção dos melhores modelos com base em métricas de avaliação. A etapa final envolveu a aplicação dos modelos de classificação selecionados em *blind wells*, que são poços sem rótulos (classes de litofácies) associados às suas variáveis (perfis geofísicos), demonstrando a eficácia dos modelos gerados. O desenvolvimento de todas as etapas mencionadas foi realizado em um ambiente Python, utilizando bibliotecas como Pandas, Numpy, Matplotlib, Seaborn, Scikit-learn e Xgboost.

Figura 7 - Fluxograma simplificado da metodologia adotada no presente estudo.



Fonte: A autora, 2023.

4.2.1 Pré-processamento dos dados

A inspeção visual foi a primeira etapa do pré-processamento, realizada para garantir que todos os dados do conjunto estivessem presentes no ambiente de programação. Após a inspeção visual, realizou-se a etapa de avaliação estatística descritiva para cada variável do conjunto de dados. Essa avaliação tem como objetivo sintetizar uma série de valores de mesma natureza, permitindo que se tenha uma visão global da variação desses valores (GUEDES et al., 2005). As medidas descritivas avaliadas incluíram a média, o desvio padrão, os valores mínimo e máximo, além das medidas separatrizes de primeiro quartil (25%),

segundo quartil (50% ou mediana) e terceiro quartil (75%). Estes representam os valores que ocupam posições no conjunto de dados dividido em quatro partes iguais.

As etapas seguintes do pré-processamento envolveram a avaliação das variáveis através da matriz de correlação e os métodos de seleção de variáveis ReliefF (KIRA; RENDELL, 1992) e InfoGain (SHANNON, 1948). O ReliefF avalia a capacidade de cada variável em distinguir instâncias com base em seus rótulos de classe e seleciona as variáveis com pontuações mais elevadas. Por outro lado, o método InfoGain calcula as informações que cada variável fornece em termos de predição de classe e leva em consideração as dependências entre as variáveis e a classe alvo, baseando-se no cálculo da entropia.

Após determinar as variáveis mais adequadas, procedeu-se com a normalização dos dados e a remoção de *outliers* (Figura 7). Os dados foram inicialmente normalizados para uma escala de 0 a 1, usando os valores mínimos e máximos de cada variável de interesse. Em seguida, identificaram-se e removeram-se outliers com base na análise do desvio padrão, excluindo-se valores além de 5 desvios padrão. Finalmente, o conjunto de dados foi dividido em conjuntos de treinamento-validação cruzada (CV) e de teste. O conjunto de treinamento-CV representou 80% e o conjunto de teste 20% dos dados originais para todos os modelos avaliados. A divisão foi realizada utilizando uma função de amostragem estratificada (*stratified*) para garantir a proporcionalidade das classes nos conjuntos de treinamento-CV e teste.

4.2.2 Processamento

A primeira etapa do processamento compreendeu o aumento artificial dos dados de treinamento (Figura 7). Essa estratégia visou mitigar o desequilíbrio de classes do conjunto de dados, no qual as classes apresentam quantidades discrepantes de amostras. Para realizar o aumento artificial, aplicamos a função SMOTE da biblioteca *Imbalanced-learn* (KOSOLWATTANA et al., 2023). Duas abordagens distintas foram empregadas para alcançar o balanceamento de classes. Na primeira abordagem, o processo de aumento artificial de amostras foi limitado ao número de amostras da classe majoritária. Na segunda abordagem, foi conduzido o aumento de até 200% das amostras das classes minoritárias, também limitando o processo até o valor de amostras da classe majoritária. Desta forma, três conjuntos de dados foram considerados para a execução dos experimentos de ML: conjunto

original (sem aumento artificial, desbalanceado); conjunto com aumento de 200%; e conjunto com aumento artificial completo (todas as classes com o número de amostra da classe majoritária).

Tendo em vista o desbalanceamento entre as classes, com objetivo de mitigar o impacto deste fator na construção do modelo de classificação, buscou-se um algoritmo capaz de lidar com conjunto de classes multiclasse desbalanceadas. Para verificar uma possível solução e selecionar o melhor algoritmo para realizar a classificação automática de litofácies complexas do Pré-sal, utilizou-se a plataforma H2O AutoML. A plataforma utiliza algoritmos e técnicas avançadas para avaliar e comparar automaticamente vários modelos em um determinado conjunto de dados. Os resultados do H2O AutoML indicaram a utilização de métodos *ensemble* para a melhor solução do problema, incluindo o algoritmo XGboost, sendo estes conhecidos por sua capacidade de lidar com grandes e complexos conjuntos de dados (CHEN; GUESTRIN, 2016).

Para ajustar o conjunto ideal de hiperparâmetros do algoritmo XGBoost ao problema de classificação proposto, adotou-se a abordagem de busca aleatória (*Random Search*), combinada com a estratégia de validação cruzada (k-*fold cross-validation*), utilizando a função *RandomizedSearchCV*. Este método de seleção envolve a amostragem aleatória de combinações de hiperparâmetros e a seleção da melhor combinação é baseada no desempenho do modelo em relação aos erros médios de validação (BERGSTRA; BENGIO, 2012). Essa é uma abordagem mais robusta nos casos em que alguns hiperparâmetros têm impacto mínimo no desempenho do modelo, permitindo a definição de um limite para o custo computacional sem que o espaço de busca dos hiperparâmetros seja restringido (BERGSTRA; BENGIO, 2012). A técnica *Random Search* foi escolhida ao invés da tradicional *Grid Search*, permitindo explorar um espaço de busca maior dentro de um tempo de processamento viável, dada a capacidade computacional disponível. Além disso, a abordagem pode mitigar uma das desvantagens do XGBoost, que é a sua suscetibilidade ao *overfitting* quando os hiperparâmetros não são adequadamente configurados (DRAHOKOUPIL, 2022).

Os hiperparâmetros utilizados para otimizar os modelos XGBoost foram o *max_depth*, *n_estimators*, *learning_rate*, *gamma*, *min_child_weight*, *subsample*, *colsample_bytree*, *colsample_bylevel*, *colsample_bynode*, *reg_alpha*, *reg_lambda*, *scale_pos_weight*. A função de cada hiperparâmetro e seus intervalos utilizados para esta busca aleatória são detalhados na Tabela 4. Para todos os experimentos realizados foi adotado como custo computacional máximo 3.000 execuções (buscas), sendo o melhor conjunto de hiperparâmetros definido pelo modelo que resultou na maior média da métrica F1-score entre todas as classes analisadas.

Após a definição dos hiperparâmetros, o algoritmo foi treinado utilizando validação cruzada estratificada com a função *StratifiedKFold*. Essa função cria dobras estratificadas que preservam a proporção de amostras de cada classe, garantindo que a validação cruzada seja realizada de forma equilibrada. Foram utilizadas cinco dobras (*n_splits*=5) para a divisão dos dados, sem embaralhamento, para garantir que, em cada iteração, uma parte fosse utilizada como conjunto de validação, enquanto as outras fossem usadas como conjunto de treinamento. Durante o treinamento-CV, o conjunto de dados foi segmentado em 5 partes iguais (k=5), denominadas *folds*, sem embaralhamento. Em cada etapa do processo, um dos *folds* foi utilizado como conjunto de validação, enquanto os quatro *folds* restantes formaram o conjunto de treinamento. Esse procedimento foi repetido 5 vezes, uma para cada *fold* (Figura 8). Esta análise oferece um maior controle sobre a etapa de treinamento-CV e apresenta uma estimativa mais confiável do erro de validação do modelo (MURPHY, 2012).

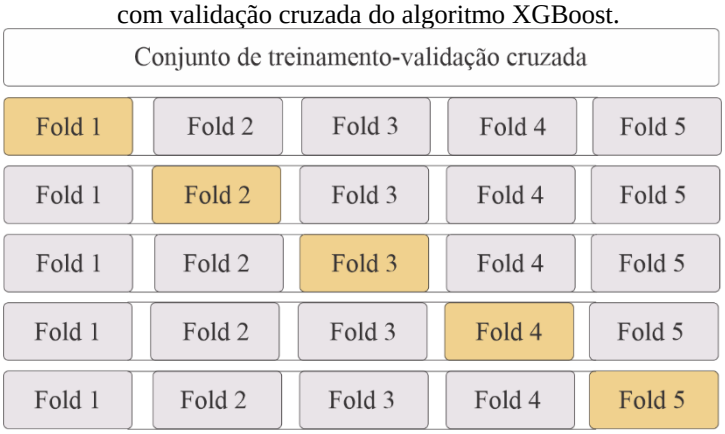
Tabela 4 - Hiperparâmetros utilizados para a otimização do algoritmo XGBoost e seus intervalos de busca.

Hiperparâmetros	Função	Intervalo de busca
max_depth[1]	Limita a profundidade máxima de uma árvore; ajuda a evitar overfitting	(1;10)[3] (1;25)[4] (1;20)[5]
n_estimators[1]	Define o número de árvores no conjunto (número de interações de treinamento)	(120;130)[3] (120;140)[4,5]
learning_rate[2]	Redução do tamanho do passo usada na atualização para evitar overfitting	(0,01; 0,11; 0,01)
gamma	Redução mínima de perdas necessária para fazer uma partição adicional em um nó da folha de uma árvore	[0]
min_child_weight[2]	Soma mínima do peso da instância (hessian) necessária em um *child*	(0,01; 0.11; 0,01)
subsample[2]	Proporção de subamostra das instâncias de treinamento, que ocorre uma vez em cada iteração de reforço	(0,1; 1,1; 0,1)
colsample_bytree[2]	Proporção de subamostra de colunas ao construir cada árvore; ocorre uma vez para cada árvore construída	(0,1; 1,1; 0,1)
colsample_bylevel[2]	Proporção de subamostra de colunas para cada nível; ocorre uma vez para cada novo nível de profundidade	(0,1; 1,1; 0,1)
colsample_bynode[2]	Proporção de subamostra de colunas para cada nó; ocorre uma vez sempre que uma nova divisão é avaliada	(0,1; 1,1; 0,1)
reg_alpha[2]	Termo de regularização L1 em pesos; quanto maior este valor, mais conservador será o modelo	(0,01; 0,11; 0,01)
reg_lambda[2]	Termo de regularização L2 em pesos; quanto maior este valor, mais conservador será o modelo	(0,01; 0,11; 0,01)
scale_pos_weight[2]	Controla o equilíbrio dos pesos positivos e negativos, útil para classes desequilibradas.	(0,01; 0,16; 0,01)

Legenda: Os intervalos dos hiperparâmetros 'max_depth' e 'n_estimators' são indicados pelos valores iniciais e finais ([1]), sendo os primeiros valores referentes aos modelos com 21 e 11 classes ([3]) e os demais para os modelos de 6 e 5 classes ([4] e [5], respectivamente). Para os demais parâmetros os intervalos são apresentados com os valores de início, fim e tamanho do incremento ([2]).
Fonte: XGBOOST, 2023. Modificado pela autora, 2023.

Após o treinamento de cada modelo, foram realizados testes com os conjuntos de dados destinados a essa etapa. A avaliação do desempenho dos modelos foi conduzida utilizando uma métrica de desempenho global (uma única para cada modelo) e também conduzida individualmente, ou seja, utilizando métricas para cada classe. As métricas acurácia, precisão, recall e F1-score foram utilizadas para a avaliação durante as etapas treinamento-CV, teste, e definição dos melhores modelos.

Figura 8 - Exemplo do método K-*fold* utilizado no treinamento com validação cruzada do algoritmo XGBoost.



Fonte: RAGB et al., 2021. Modificada pela autora, 2023.

Cada uma das métricas de avaliação utilizadas mede uma característica específica ou a relação entre as distintas características de desempenho dos modelos. A métrica de avaliação acurácia mede o percentual geral de previsões corretas do modelo (Equação 1). A precisão avalia a proporção de predições positivas corretas em relação ao total de predições positivas (Equação 2). O recall mede a proporção de verdadeiros positivos em relação ao total de casos positivos verdadeiros e negativos falsos (Equação 3). O F1-score é uma métrica que considera tanto a precisão quanto o recall, fornecendo uma medida equilibrada do desempenho do modelo (Equação 4).

$$Acurácia = (TP + TN)/(TP + TN + FP + FN)$$

(1)

$$Precisão = TP/(TP + FP)$$

(2)

$$Recall = TP/(TP + FN)$$

(3)

$$F1\ score = 2 \times (\frac{precisão \times recall}{precisão + recall})$$

(4)

Onde: TP e TN indicam o número de amostras corretamente classificadas (*true positive* ou *true negative*), e; FP e FN representam as amostras erroneamente classificadas (*false positive* ou *false negative*).

Por fim, após a definição dos melhores modelos, realizou-se a avaliação através da matriz de confusão. Esta compara as classes reais com as previstas, indicando quantas vezes o modelo acertou, errou e como os erros foram distribuídos entre as outras classes. Em uma matriz de confusão, as classificações corretas são dispostas ao longo da diagonal, enquanto todas as outras células mostram classificações incorretas (VISA et al., 2011). Posteriormente, aplicou-se esses modelos aos poços sem dados de classificação conhecidos, *blind wells*.

5 **RESULTADOS E DISCUSSÕES**

Nas seções subsequentes, são apresentados os resultados e discussões relacionados à abordagem de ML supervisionado e algoritmo XGBoost. Este algoritmo foi empregado na classificação de litofácies que ocorrem nos reservatórios da FBVE no Campo de Tupi, localizado na Bacia de Santos.

5.1 **Análises na base de dados utilizada**

O conjunto de dados utilizado para a aplicação do ML supervisionado na classificação automatizada de litofácies é composto por dados de seis poços. Os dois poços com informações de litofácies (proveniente de testemunhos) e com dados de perfilagem convencional e avançada foram utilizados para o desenvolvimento dos modelos de classificação automatizada. Os outros quatro poços foram utilizados como *blind wells*, selecionados por apresentarem os mesmos dados de perfilagem que os dois poços com informações de litofácies (Tabela 1).

Os dois poços utilizados no treinamento-CV e teste apresentaram 21 tipos de litofácies (Tabela 2) e 21 perfis geofísicos (Tabela 3). Esses poços contêm um total de 1.778 amostras de litofácies e 37.338 exemplos de perfis. No processo, os dados de litofácies foram empregados como rótulos, enquanto os perfis geofísicos serviram como variáveis para o desenvolvimento dos modelos classificadores. Além disso, os poços designados como *blind wells* apresentam uma espessura total de 922 m de formação, correspondendo a 6.052 exemplos de perfis geofísicos (um exemplo a cada 0,1524 m), que foram empregados na aplicação do modelo de ML supervisionado.

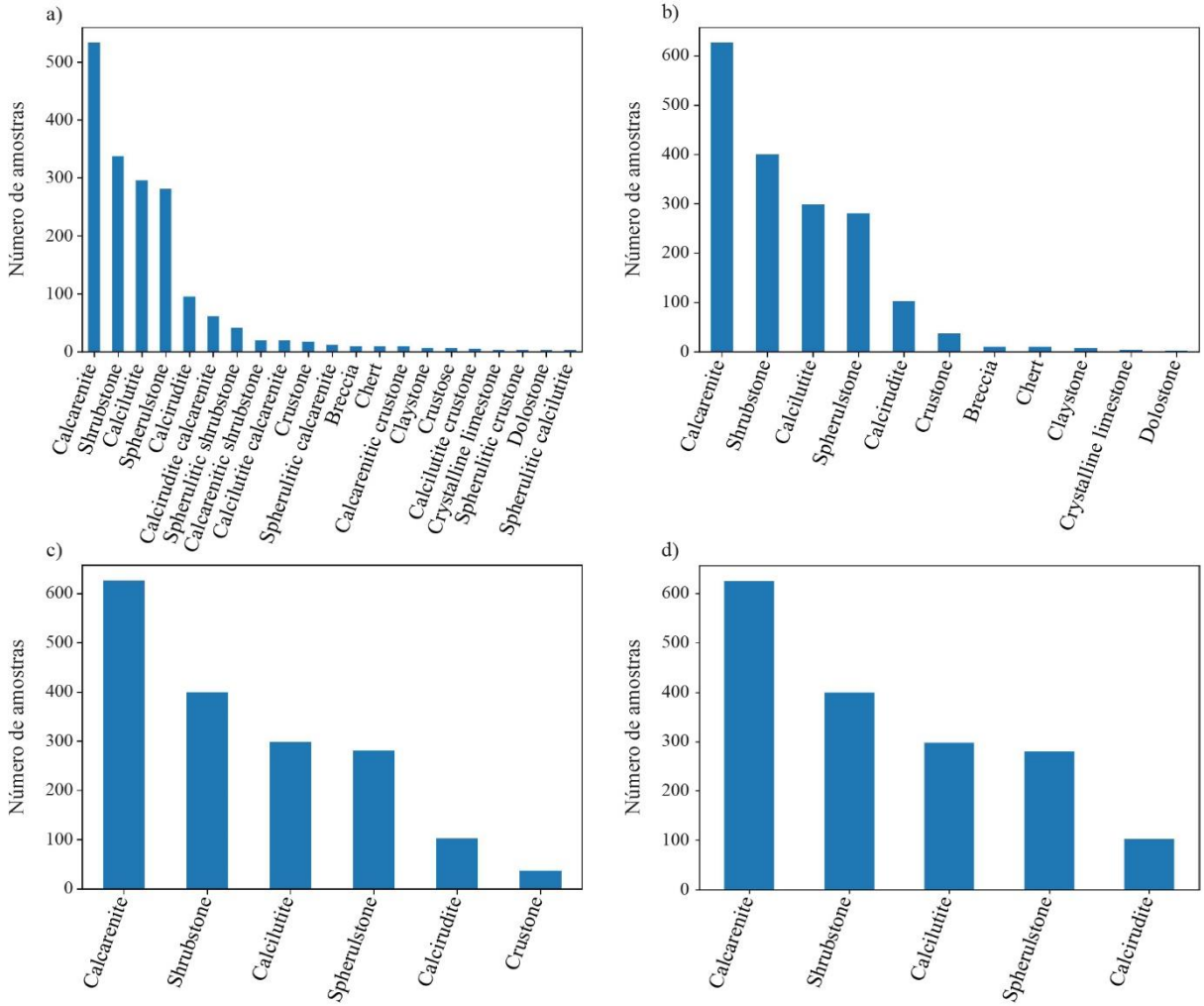
A análise preliminar do conjunto de dados revelou que ele é caracterizado como um conjunto de dados multiclasse e desbalanceado. A característica multiclasse indica que o conjunto de dados possui mais de duas classes para realizar a tarefa de classificação automatizada, enquanto o desbalanceamento de classes indica que a classe majoritária possui uma quantidade superior de amostras em relação às classes minoritárias. Dentre as 21 classes de litofácies, a classe majoritária é composta por mais de 500 amostras, enquanto as classes intermediárias têm variações em torno de 100 amostras. Em contraste, cada uma das classes

minoritárias contém menos de 30 amostras, com algumas tendo apenas 3 amostras, representando menos de 1% do total de amostras (Tabela 2; Figura 9).

O desbalanceamento de classes em um conjunto de dados multiclasse é uma questão que causa grande impacto nos resultados, pois sistemas de aprendizado podem enfrentar desafios ao realizar a classificação devido a desigualdades entre as classes ou à sobreposição delas (BATISTA; PRATI; MONARD, 2004, 2005; PRATI; BATISTA; MONARD, 2004; GARCÍA; SÁNCHEZ; MOLLINEDA, 2012). Especificamente, em modelos nos quais uma classe apresenta um número significativamente maior de exemplos do que a outra, isso tende a comprometer a eficácia da classificação da classe minoritária (BATISTA; PRATI; MONARD, 2004).

Devido ao provável impacto negativo do desbalanceamento de classes, e definição de que o algoritmo adequado para o desenvolvimento do estudo seria o XGBoost, aplicaram-se técnicas de agrupamento e de associação de classes de litofácies como alternativas para mitigar o impacto do desbalanceamento de classes do conjunto de dados. Uma das abordagens testou o algoritmo de agrupamento não supervisionado *K-means* (LIKAS; VLASSIS; J. VERBEEK, 2003; WU, 2008), o qual resultou em desempenho insatisfatórios no agrupamento das litofácies da FBVE, no Campo de Tupi. Dessa forma, adotou-se uma abordagem que associou litofácies geologicamente semelhantes em grupos, incluindo algumas das classes menos representativas nesses grupos. Portanto, além do conjunto de dados original, que continha 21 classes de litofácies, outros três conjuntos de dados com variações no número de classes e amostras foram utilizados nas etapas subsequentes de construção dos modelos de classificação (Figura 9).

Figura 9 - Histogramas dos conjuntos de dados utilizados.



Legenda: (a) conjunto com 21 classes de litofácies; (b) 11 classes; (c) 6 classes; (d) 5 classes.
Fonte: A autora, 2023.

Em relação aos dados utilizados nos experimentos de ML, foram adotadas quatro abordagens. A primeira abordagem utilizou as 21 classes de litofácies para criar o modelo de 21 classes (Figura 9a; Tabela 2). A segunda, associou as classes em 11 grupos de litofácies (Figura 9b; Tabela 5). A terceira, baseada nos 11 grupos, eliminou 5 classes minoritárias, cada uma com menos de 37 amostras, resultando em 6 grupos utilizados no modelo de 6 classes (Figura 9c; Tabela 5). As 5 classes de litofácies eliminadas do modelo foram *breccia*, *chert*, *claystone*, *crystaline-limestone* e *dolostone*, devido à sua falta de similaridade geológica com os grupos de litofácies existentes. Na última abordagem, utilizou-se as classes majoritárias com mais de 100 amostras, formando 5 grupos de litofácies, modelo de 5 classes (Figura 9d; Tabela 5), resultando na eliminação da classe *crustone* desse conjunto de dados.

Tabela 5 - Litofácies, grupos de litofácies e suas respectivas quantidades de amostras.

Litofácies	Grupo de litofácies	Amostras (n)
Breccia	*Breccia*	10
Chert	*Chert*	10
Claystone	*Claystone*	7
Crustone *Calcarenitic crustone* *Spherulitic crustone* *Calcilutite crustone*	*Crustone*	37
Crystalline limestone	*Crystalline limestone*	4
Dolostone	*Dolostone*	3
Calcarenite *Calcirudite calcarenite* *Calcilutite calcarenite* *Spherulitic calcarenite*	*Calcarenite*	626
Calcilutite *Spherulitic calcilutite*	*Calcilutite*	298
Calcirudite *Crustose*	*Calcirudite*	102
Shrubstone *Calcarenitic shrubstone* *Spherulitic shrubstone*	*Shrubstone*	400
Spherulstone	*Spherulstone*	281

Legenda: n- número de amostras.
Fonte: A autora, 2023.

A etapa seguinte consistiu na aplicação dos métodos de seleção de variáveis ReliefF (KIRA; RENDELL, 1992) e InfoGain (SHANNON, 1948), além da avaliação da matriz de correlação de Pearson entre as variáveis. Este processo revelou baixo desempenho de seis variáveis (perfis geofísicos): porosidade neutrônica, resistividade rasa e média, NMR - média logarítmica, volume de fluido livre e fração volumétrica de fluido, levando à remoção destes do conjunto de dados. Considerando o conhecimento geológico, esses perfis foram eliminados por não apresentarem boa correlação com as litofácies carbonáticas e por estarem predominantemente relacionados ao fluido de perfuração ou à sua interação com a formação. Portanto, esse procedimento resultou na utilização de 15 variáveis para a construção dos

modelos de classificação, apresentadas na Tabela 6, juntamente com algumas de suas estatísticas descritivas.

Ao analisar os dados das estatísticas apresentadas na Tabela 6, destaca-se que, se compararmos o resultado do desvio padrão com a média dos dados, AL_WF, TI_WF, IRON_WF e SU_WF são as variáveis cujos conjuntos tiveram maior variação, sendo a última apresenta essa relação superior a 300%. Desta forma, tem-se que os dados dessas variáveis são bastante heterogêneos entre si, provavelmente relacionado à presença de litofácies com características distintas no conjunto de dados. Por outro lado, RHOB e DTC apresentaram coeficiente de variação inferior a 10%, sendo que uma baixa variabilidade de uma variável indica que esta possui um comportamento mais homogêneo para todo o conjunto de classes analisados, o que foi o caso do perfil de densidade, por exemplo.

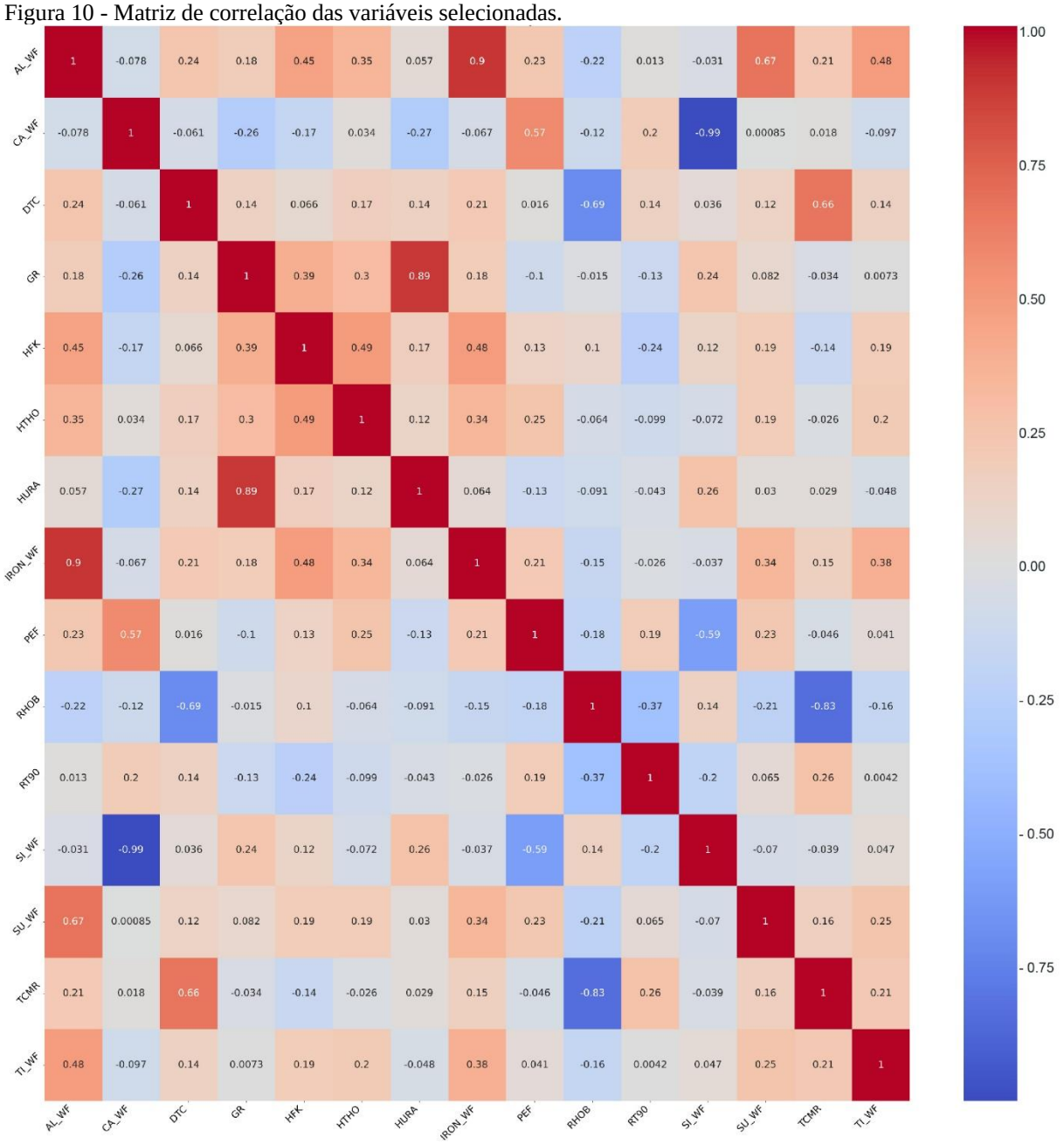
Além de entender o comportamento de cada variável isoladamente, a matriz de correlação (Figura 10) indicou a relação entre as variáveis do conjunto de dados (Tabela 6). Os testes de correlação são técnicas empregadas para investigar associações entre o comportamento de grupos de variáveis, facilitando a construção de modelos hipotéticos que devem então ser confirmados por meio de experimentos dedicados (MIOT, 2017). Os valores mais próximos a 1 (um) indicam que as variáveis estão fortemente correlacionadas positivamente. Os valores mais próximos a -1 indicam que as variáveis estão fortemente correlacionadas negativamente, e os valores próximos a 0 indicam que há uma correlação fraca ou inexistente entre as variáveis. A correlação positiva indica que as variáveis têm relação linear direta e tendem a aumentar ou diminuir juntas. A correlação negativa indica que as variáveis têm relação inversa, ou seja, quando uma delas aumenta, a outra diminui. Os valores próximos a 0 indicam que as variáveis são independentes entre si.

Tabela 6 - Variáveis selecionadas para o desenvolvimento do ML supervisionado.

Perfis geofísicos	Mnemônicos	Média	DP	Mín.-Máx.	Q1	Q2	Q3
Raios gama	GR	22,82	8,71	7,55-66,81	16,51	21,42	26,47
Densidade	RHOB	2,60	0,07	2,20-2,85	2,57	2,61	2,65
Compressional slowness	DTC	58,76	4,71	51,12-77,83	55,43	57,52	61,35
Resistividade profunda	RT-90	267,27	284,18	3,18-1833,11	75,54	170,38	352,67
GR espectral – concent. de potássio	HFK	0,0017	0,001	0,0005-0,0122	0,0011	0,0014	0,0020
GR espectral – concent. de urânio	HURA	1,81	0,95	0,42-6,40	1,09	1,64	2,19
GR espectral – concent. de tório	HTHO	1,45	0,41	0,43-3,02	1,17	1,39	1,69
Fator fotoelétrico	PEF	4,04	0,55	2,78-6,98	3,71	4,07	4,39
ECS - fração de peso de alumínio	AL_WF	2,42e-3	3,64e-3	0-4,44e-2	3,35e-7	1,09e-3	3,42e-3
ECS - fração de peso de cálcio	CA_WF	0,3036	0,0359	0,18-0,382	0,28	0,31	0,33
ECS - fração de peso de ferro	IRON_WF	0,0008	0,0015	0-0,0146	0	0	0,0009
ECS - fração de peso de silício	SI_WF	1,09e-1	4,20e-2	0-2,59e-1	7,79e-2	1,01e-1	1,35e-1
ECS - fração de peso de enxofre	SU_WF	0,0034	0,0114	0-0,2360	0	0,0014	0,0046
ECS - fração de peso de titânio	TI_WF	2,57e-4	4,37e-4	0-4,02e-4	0	5,27e-7	3,85e-3
NMR - porosidade total	TCMR	0,09	0,04	0,04-0,39	0,07	0,09	0,11

Legenda: DP- Desvio Padrão; Mín-Máx- Valores mínimos e máximos das variáveis; Q1- primeiro quartil (25%); Q2- segundo quartil ou mediana (50%); Q3- terceiro quartil (75%).
Fonte: A autora, 2023.

Figura 10 - Matriz de correlação das variáveis selecionadas.



Legenda: AL_WF- ECS - fração de peso de alumínio; CA_WF- ECS - fração de peso de cálcio; DTC- *Compressional slowness*; GR- Raios gama; HFK- GR espectral - concentração de potássio; HTHO- GR espectral - concentração de tório; HURA- GR espectral - concentração de urânio; IRON_WF- ECS - fração de peso de ferro; PEF- Fator fotoelétrico; RHOB- Densidade; RT90- Resistividade profunda; SI_WF- ECS - fração de peso de silício; SU_WF- ECS - fração de peso de enxofre; TCMR- NMR - porosidade total, e; TI_WF- ECS - fração de peso de titânio.

Fonte: A autora, 2023.

Na Figura 10, é possível observar que algumas variáveis possuem correlação positiva, algumas não apresentam correlação linear, e que outras possuem correlação negativa entre si. As variáveis que apresentaram as maiores correlações positivas foram: AL_WF e IRON_WF (0,90), GR e HURA (0,89), SU_WF e AL_WF (0,67), DTC e TCMR (0,66). As variáveis que

apresentaram as maiores correlações negativas foram: CA_WF e SI_WF (-0,99), RHOB e TCMR (-0,83), e RHOB e DTC (-0,69). Considerou-se a remoção de variáveis altamente correlacionáveis entre si para evitar multicolinearidade, que poderia afetar negativamente a construção dos modelos. Contudo, a possibilidade de remoção dessas variáveis foi avaliada com cautela, pois a análise não forneceu evidências de dependência direta ou de causalidade entre elas, revelando apenas que tendem a variar em conjunto (MIOT, 2017). Dessa forma, após a análise da matriz de correlação, nenhuma variável foi removida e o desenvolvimento do estudo prosseguiu com as 15 variáveis selecionadas (Tabela 6).

5.2 **Machine learning supervisionado - Classificação de litofácies**

Após a seleção e avaliação das variáveis que fariam parte do conjunto de dados que seria explorado na etapa de ML supervisionado, as etapas de pré-processamento dos dados foram realizadas, incluindo normalização e remoção de outliers. A normalização foi empregada para padronizar a escala das variáveis, possibilitando uma análise mais equitativa pelos modelos. Essa etapa se mostrou necessária devido à significativa diferença entre as escalas das diferentes variáveis (JAMAL et al., 2014). Com isso, todas as variáveis foram ajustadas para um intervalo de 0 a 1.

Quanto à remoção de outliers, esta foi aplicada para excluir do conjunto de dados aqueles valores que estavam fora do desvio padrão, considerando-se, neste caso, valores acima de cinco desvios padrão. O limite de cinco desvios padrão foi escolhido considerando que os perfis provêm de diferentes litofácies, as quais possuem características distintas entre si. Portanto, a adoção de um critério mais amplo para a classificação de outliers poderia resultar na exclusão de valores plausíveis, que se diferenciam dos demais por pertencerem a uma litofácies diferente. Após esse processo, o número de amostras remanescentes para o desenvolvimento dos modelos de classificação foi reduzido para 1.744.

Posteriormente, o conjunto foi dividido em duas partes: uma destinada ao treinamento e validação cruzada (CV), correspondendo a 80% dos dados, e a outra para teste, compreendendo os 20% restantes. Esta divisão foi feita de forma estratificada, utilizando a função *stratified*, para assegurar uma representação adequada de todas as classes nos conjuntos de dados e manter a proporção entre as classes. Esse procedimento garantiu a representatividade das classes e das amostras em todas as etapas do experimento. A Tabela 7

ilustra um exemplo da aplicação dessa função no conjunto de dados, que neste caso é composto por seis classes.

Tabela 7 - Resultado da divisão do conjunto de dados composto por seis classes de litofácies.

Etapa	*Calcarenite*	*Shrubstone*	*Calcilutite*	*Spherulstone*	*Calcirudite*	*Crustone*
Treinamento-CV	489	317	237	218	78	28
Teste	123	79	60	54	19	7

Fonte: A autora, 2023.

Entretanto, apesar da estratégia de divisão estratificada, o desafio do desbalanceamento de classes ainda persistia, o que levou à implementação da técnica SMOTE (*Synthetic Minority Over-sampling Technique*) como um passo adicional ao processamento dos dados de treinamento-CV, enquanto que no conjunto de teste foram mantidos os dados originais. Esta técnica gera amostras sintéticas das classes minoritárias, aumentando seu tamanho e equilibrando a distribuição das classes (KOSOLWATTANA et al., 2023). O SMOTE seleciona uma amostra da classe minoritária e os seus k vizinhos mais próximos (*K-neighbors*), criando novas amostras através da interpolação entre a amostra selecionada e seus vizinhos (CHAWLA et al., 2002; HE; GARCIA, 2009). No caso dos modelos de 21 e 11 classes, devido ao número reduzido de amostras em algumas classes, o aumento artificial de amostras foi realizado com K-neighbors igual a 1. Já para os modelos de 6 e 5 classes, com uma quantidade maior de amostras ($n \geq 37$) nas classes, o aumento foi efetuado com K-neighbors igual a 5.

Foram utilizadas duas estratégias de aumento de amostras, gerando dois conjuntos de dados distintos para cada um dos conjuntos originais (compostos por 21, 11, 6 e 5 classes). Na primeira estratégia, empregou-se a técnica SMOTE, limitando o número de amostras de todas as classes ao da classe majoritária. A segunda estratégia envolveu o aumento das amostras de cada classe em até 200%, sempre respeitando o limite imposto pelo número de amostras da classe majoritária. Os resultados dessas estratégias de aumento artificial de amostras são apresentados na Tabela 8, que detalha as quantidades de amostras para treinamento, validação cruzada e teste de cada modelo.

Tabela 8 - Composição dos modelos utilizados para treinamento, validação cruzada e teste, incluindo a porcentagem de aumento artificial de dados aplicada a cada modelo.

Modelo	Classes	Total de amostras	Amostras Treinamento-CV	% Aumento artificial	Amostras teste
1		1.744	1.395	0	
2	21	9.085	8.736	=CM	349
3		2.784	2.435	≤200	
4		1.744	1.395	0	
5	11	5.728	5.368	=CM	348
6		2.704	2.355	≤200	
7		1.709	1.367	0	
8	6	3.276	2.928	=CM	341
9		2.697	2.355	≤200	
10		1.675	1.340	0	
11	5	2.785	2.440	=CM	335
12		2.526	2.191	≤200	

Legenda: Treinamento-CV- treinamento com validação cruzada; CM- aumento artificial das amostras de todas as classes minoritárias até o valor máximo (número de amostras da classe majoritária).
Fonte: A autora, 2023.

Considerando que o algoritmo XGBoost requer hiperparâmetros específicos para a construção de cada modelo de classificação, a definição desses hiperparâmetros, após o processo de balanceamento de dados, foi realizada utilizando o método *Random search* com validação cruzada. A busca pelos melhores hiperparâmetros do algoritmo, através do ajuste específico de hiperparâmetros para cada modelo, impactou positivamente a classificação das litofácies da FBVE. Ajustes foram realizados individualmente nos hiperparâmetros de cada modelo para otimizar seu desempenho na classificação das litofácies. Essas variações refletem a adaptação específica dos hiperparâmetros a cada contexto de mode (Tabela 9). O *max_depth* variou de 9 a 14, indicando a profundidade máxima das árvores nos modelos. O *n_estimators* oscilou entre 120 e 129, enquanto o *learning_rate* esteve entre 0,07 e 0,1. O *gamma* foi mantido constante em 0 para todos os modelos. O *min_child_weight* variou de 0,01 a 0,07. Os valores de *subsample* ficaram entre 0,6 e 1,0, e o *colsample_bytree* oscilou de 0,8 a 1,0. Já o *colsample_bylevel* e *colsample_bynode* variaram entre 0,3 e 1,0, mostrando uma variação significativa na subamostragem por nível de árvore e por nó. Os parâmetros *reg_alpha* e *reg_lambda* variaram de 0,01 a 0,1, indicando diferenças nas regularizações L1 (*alpha*) e L2 (*lambda*) aplicadas nos modelos.

Tabela 9 - Resultado da busca dos melhores hiperparâmetros para cada modelo selecionado.

Hiperparâmetros	Modelo 2	Modelo 5	Modelo 9	Modelo 11
max_depth	9	9	10	14
n_estimators	121	129	128	120
learning_rate	0,09	0,1	0,1	0,07
gamma	0	0	0	0
min_child_weight	0,01	0,07	0,04	0,04
subsample	1,0	0,7	0,6	0,7
colsample_bytree	0,8	1,0	0,9	0,8
colsample_bylevel	0,6	0,9	0,6	1,0
colsample_bynode	0,5	0,3	0,8	0,4
reg_alpha	0,05	0,1	0,01	0,01
reg_lambda	0,1	0,07	0,07	0,06
scale_pos_weight	0,05	0,15	0,09	0,1

Fonte: A autora, 2023.

Após a definição dos parâmetros mais adequados para cada modelo, foi realizado o treinamento dos 12 modelos utilizando validação cruzada, com K-*fold* igual a 5. Os resultados das métricas de avaliação acurácia, precisão, recall e F1-score para cada modelo quando aplicado ao conjunto de validação são exibidos na Tabela 10. Nota-se nos resultados apresentados que, com exceção dos modelos cujo conjunto de dados não passou pelo aumento artificial de amostras, todos os outros modelos apresentaram todas as métricas superiores a 80%, indicando um ótimo desempenho na etapa de treinamento-CV.

Os modelos treinados e validados foram então aplicados ao conjunto de teste e, a partir da análise das métricas de avaliação de modelos aplicados a este conjunto, apresentadas na Tabela 11, foram selecionados os melhores modelos de cada conjunto de dados para a predição de litofácies. Os modelos com as melhores métricas foram o modelo 2, formado por 21 classes, e o modelo 5, com 11 classes de litofácies, ambos caracterizados pelo aumento artificial total de amostras. Além destes, o modelo 9, formado por 6 classes com um aumento artificial de amostras de 200%, e o modelo 11, composto por 5 classes de litofácies com aumento artificial total de amostras, também se destacaram (Figura 11a-d).

Tabela 10 - Métricas de avaliação dos 12 modelos gerados aplicados ao conjunto de CV.

Modelo	Classes	% aumento	Acurácia	Precisão	Recall	F1-score
1		0	0,62	0,50	0,32	0,36
2	21	=CM	0,96	0,96	0,96	0,96
3		≤200	0,85	0,94	0,82	0,87
4		0	0,43	0,18	0,14	0,12
5	11	=CM	0,89	0,89	0,89	0,89
6		≤200	0,86	0,89	0,81	0,84
7		0	0,67	0,70	0,57	0,61
8	6	=CM	0,88	0,88	0,88	0,88
9		≤200	0,83	0,84	0,85	0,84
10		0	0,67	0,68	0,59	0,62
11	5	=CM	0,85	0,85	0,85	0,85
12		≤200	0,83	0,84	0,83	0,84

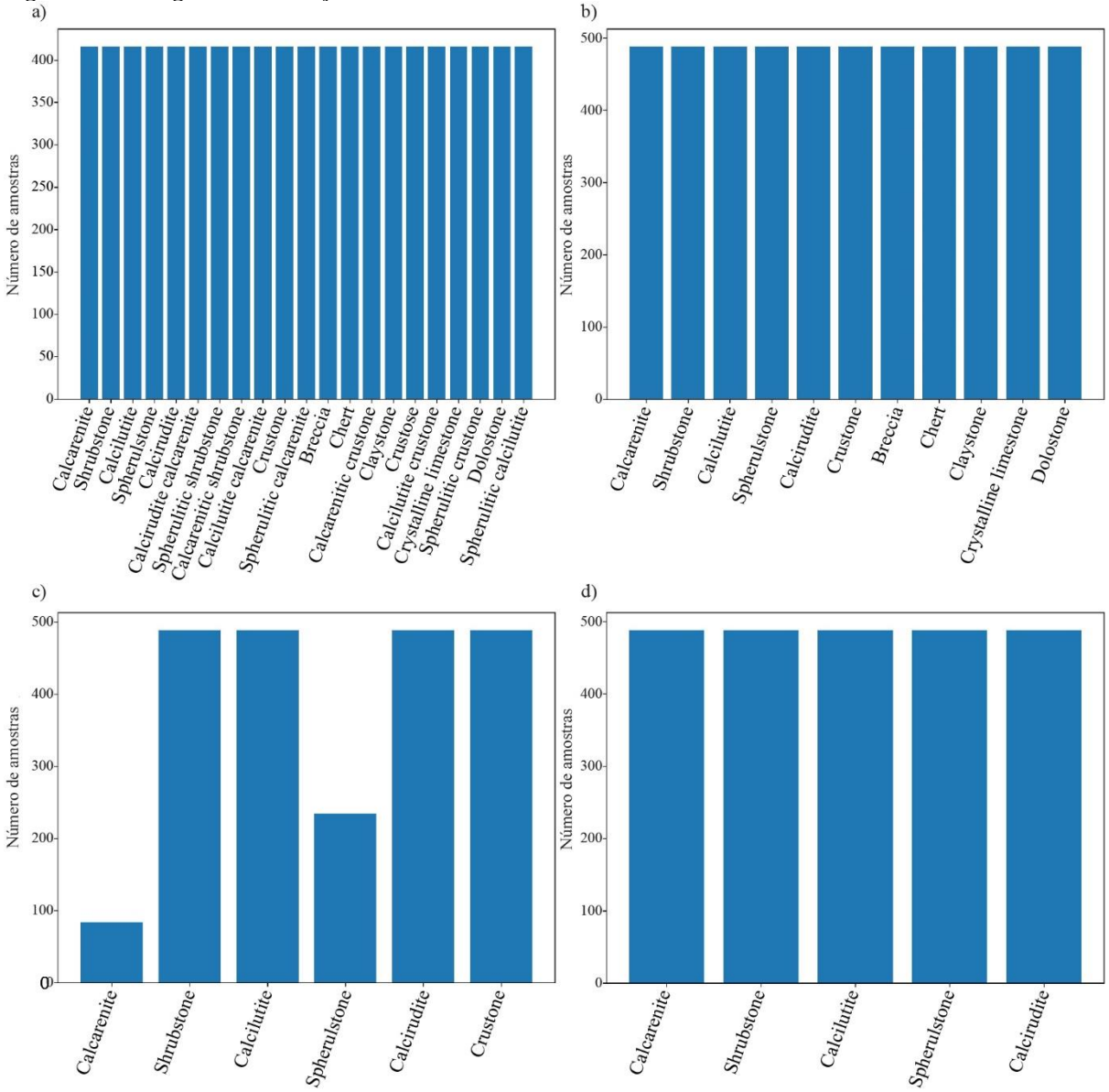
Legenda: CM- classe majoritária.
Fonte: A autora, 2023.

Tabela 11 - Métricas de avaliação dos 12 modelos gerados aplicados ao conjunto de teste.

Modelo	Classes	% aumento	Acurácia	Precisão	Recall	F1-score	n classes
1		0	0,63	0,61	0,63	0,60	15
2	21	=CM	0,69	0,69	0,69	0,68	17
3		≤200	0,68	0,69	0,68	0,67	17
4		0	0,62	0,61	0,61	0,61	7
5	11	=CM	0,64	0,63	0,63	0,63	8
6		≤200	0,63	0,62	0,63	0,62	7
7		0	0,69	0,70	0,69	0,68	6
8	6	=CM	0,70	0,71	0,70	0,70	6
9		≤200	0,72	0,73	0,72	0,72	6
10		0	0,65	0,68	0,65	0,65	5
11	5	=CM	0,70	0,71	0,70	0,70	5
12		≤200	0,68	0,68	0,68	0,68	5

Legenda: CM- classe majoritária; n classes- quantidade de classes corretamente identificadas.
Fonte: A autora, 2023.

Figura 11 - Histogramas dos conjuntos de dados de treinamento-CV dos melhores modelos obtidos.



Legenda: (a) modelo 2: conjunto de 21 classes de litofácies com aumento artificial CM; (b) modelo 5: conjunto de 11 classes de litofácies com aumento artificial CM; (c) modelo 9: conjunto de 6 classes de litofácies com aumento artificial de 200%; (d) conjunto de 5 classes de litofácies com aumento artificial CM; CM- aumento artificial de amostras ≤ classe majoritária.

Fonte: A autora, 2023.

O modelo 2 alcançou uma acurácia de 0,69, precisão de 0,69, F1-score de 0,68 e recall de 0,69, classificando corretamente 17 das 21 classes de litofácies. O modelo 5 registrou uma acurácia de 0,63, precisão de 0,62, F1-score de 0,63 e recall de 0,63, com a classificação correta de 8 das 11 classes de litofácies. Já o modelo 9 apresentou uma acurácia de 0,72, precisão de 0,72, F1-score de 0,71 e recall de 0,72, acertando todas as 6 classes de litofácies. Por fim, o modelo 11 teve uma acurácia de 0,70, precisão de 0,71, F1-score de 0,70 e recall de 0,70, classificando corretamente todas as 5 classes de litofácies (Tabela 11).

A capacidade de cada modelo selecionado para classificar corretamente um número significativo de classes de litofácies valida a eficácia das técnicas de pré-processamento e balanceamento de dados adotadas, tendo em vista que os quatro modelos com melhor performance foram submetidos a um aumento artificial de amostras, seja total ou de até 200%, limitado à quantidade de amostras da classe majoritária. Essa condição indica que a ampliação do volume de dados desempenhou um papel significativo no aprimoramento do desempenho dos modelos de classificação. Esse resultado corrobora com os apresentados por Parsa (2021) e Ibrahim et al. (2023), que validaram a abordagem de utilização do algoritmo XGboost aliado a técnicas de aumento de dados como sendo uma alternativa para melhorar a eficiência de previsão dos modelos quando utilizados conjuntos de dados desequilibrados, principalmente na classificação das classes minoritárias (IBRAHIM et al., 2023).

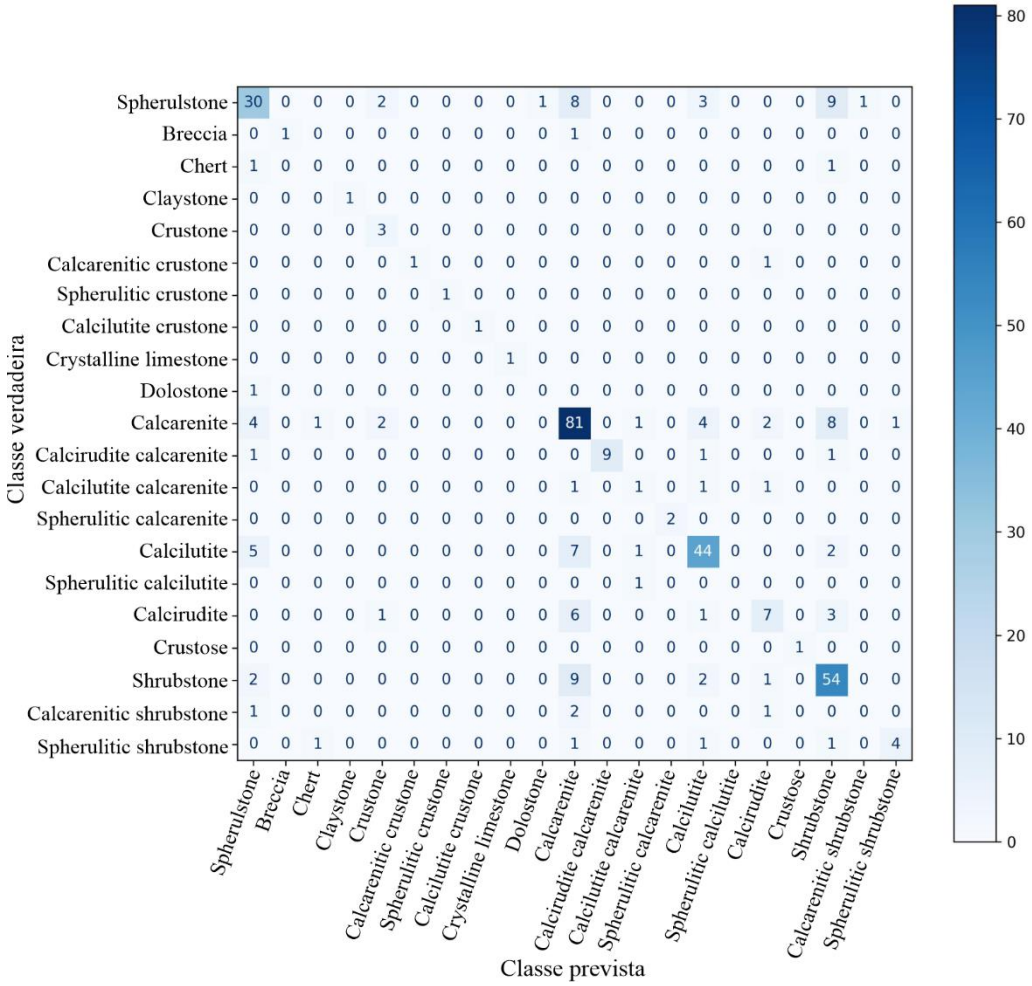
Com o objetivo de avaliar o desempenho dos modelos selecionados para cada uma das classes de litofácies, são apresentados nas Figura 12 a Figura 15 os resultados das classificações em forma de matriz de confusão, permitindo a comparação entre a classe predita e a verdadeira. Esta visualização é importante pois permite observar se os erros de classificação ocorreram em determinados tipos de litofácies específicos e entre quais classes ocorreram.

A avaliação da matriz de confusão do modelo 2 (Figura 12) demonstrou que 17 das 21 classes foram corretamente identificadas em pelo menos uma predição, sendo as não identificadas *chert, dolostone, spherulitic-calcilutite* e *calcarenitic-shrubstone*. Com exceção da classe *calcarenitic-shrubstone,* as demais cujos dados foram classificados erroneamente destacam-se por estarem no grupo de litofácies com menor quantidade de amostras (igual ou inferior a 10 no total). Das classes identificadas corretamente, 7 tiveram um aproveitamento de 100% de acertos, incluindo *claystone, crustone, spherulitic-crustone, calcilutite-crustone, crystalline-limestone, spherulitic-calcarenite* e *crustose*. As demais classes apresentaram desempenho variável, com destaque para as quatro classes majoritárias, que variaram de 55% (*spherulstone*) a 79% (*shrubstone*) de acerto em suas classificações.

Destaca-se também que o modelo de classificação 2 (Figura 12) apresentou diferentes taxas de erro ao confundir amostras de 14 classes de litofácies. A classe *spherulstone* foi incorretamente classificada como *shrubstone* e *calcarenite*. Observou-se que uma amostra da classe *breccia* foi erroneamente prevista como *calcarenite* e as amostras de *Chert* como *spherulstone* e *shrubstone*. Da mesma forma, uma amostra de *calcarenitic-crustone* foi classificada incorretamente como *calcirudite,* e a amostra de *dolostone* como *spherulstone*. O erro predominante na classe *calcarenite* foi a classificação errônea como *shrubstone*. Além

disso, amostras de *calcirudite-calcarenite* foram identificadas como *spherulstone*, *calcarenite* e *shrubstone*, e amostras de *calcilutite-calcarenite* foram confundidas com *calcarenite*, *calcilutite* e *calcirudite*. Erros na classe *calcilutite* incluíram classificações incorretas como *calcarenite* e *spherulstone*. A amostra de *spherulitic-calcilutite* foi classificada como *calcilutite-calcarenite* e as amostras de *calcirudite* foram erroneamente classificadas como *calcarenite*, *shrubstone* e *crustone*. A classe *shrubstone* apresentou a maior confusão associada à classe *calcarenite*, e as amostras de *calcarenitic-shrubstone* foram previstas incorretamente como *spherulstone*, *calcarenite* e *calcirudite*. Por fim, amostras de *spherulitic-shrubstone* foram classificadas como *chert*, *calcarenite* e *shrubstone*. Tendo em vista os resultados obtidos da aplicação do modelo 2, tem-se que o seu desempenho, mesmo com o aumento artificial de dados, pode ter sido afetado pela ausência de diversidade de informações em algumas classes com pouquíssimas amostras reais utilizadas no conjunto de treinamento.

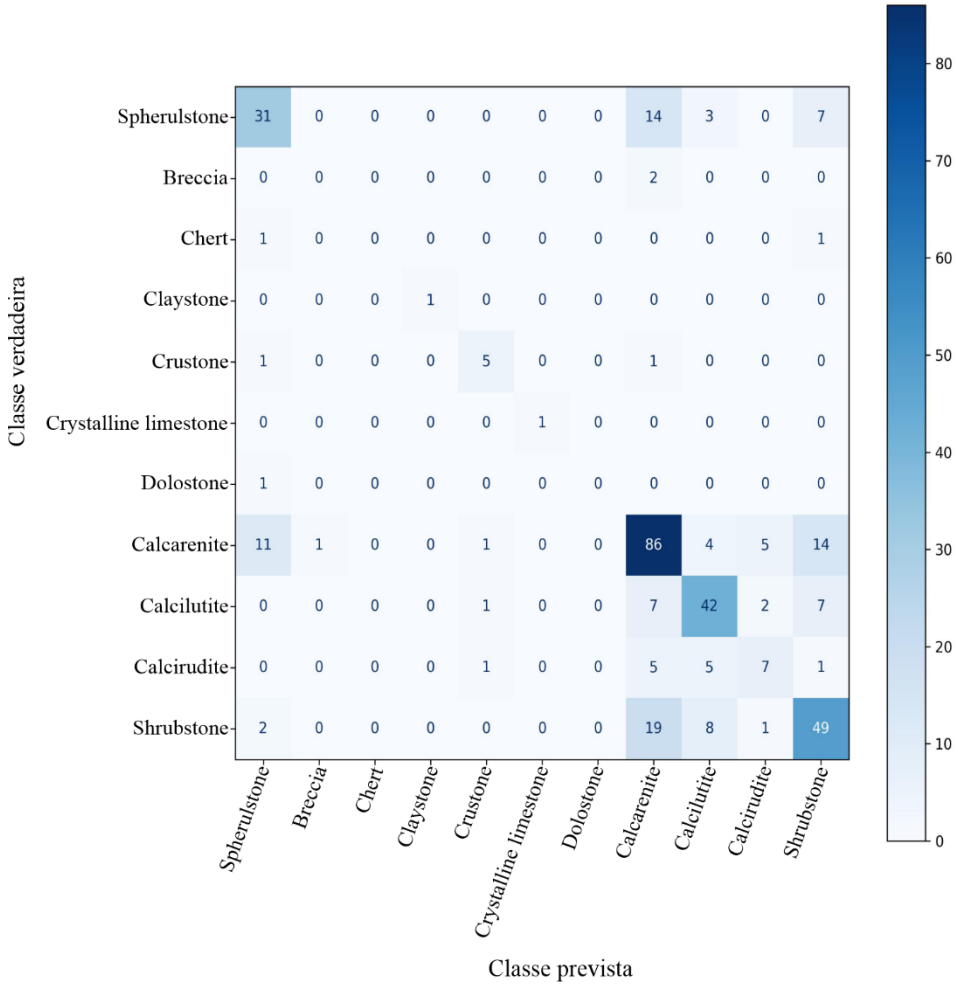
Figura 12 - Matriz de confusão do modelo 2, composto por 21 classes de litofácies, aplicado ao conjunto de teste.



Fonte: A autora, 2023.

Já a matriz de confusão do modelo 5 (Figura 13) mostrou uma maior variação nos resultados de predição por classe. As classes com desempenho mais destacado foram *claystone* e *crystalline limestone*, ambas com uma taxa de acerto de 100%. As classes *calcarenite* e *crustone* apresentaram desempenho acima de 71%. Desempenhos ligeiramente inferiores foram observados para *calcilutite*, com 57,53%, *shrubstone*, com 62,03%, e *spherulstone*, com 65,96%. *calcirudite* registrou um desempenho menos satisfatório, com 38,89%. As classes *breccia*, *chert* e *dolostone* não tiveram acertos.

Figura 13 - Matriz de confusão do modelo 5, composto por 11 classes de litofácies, aplicado ao conjunto de teste.



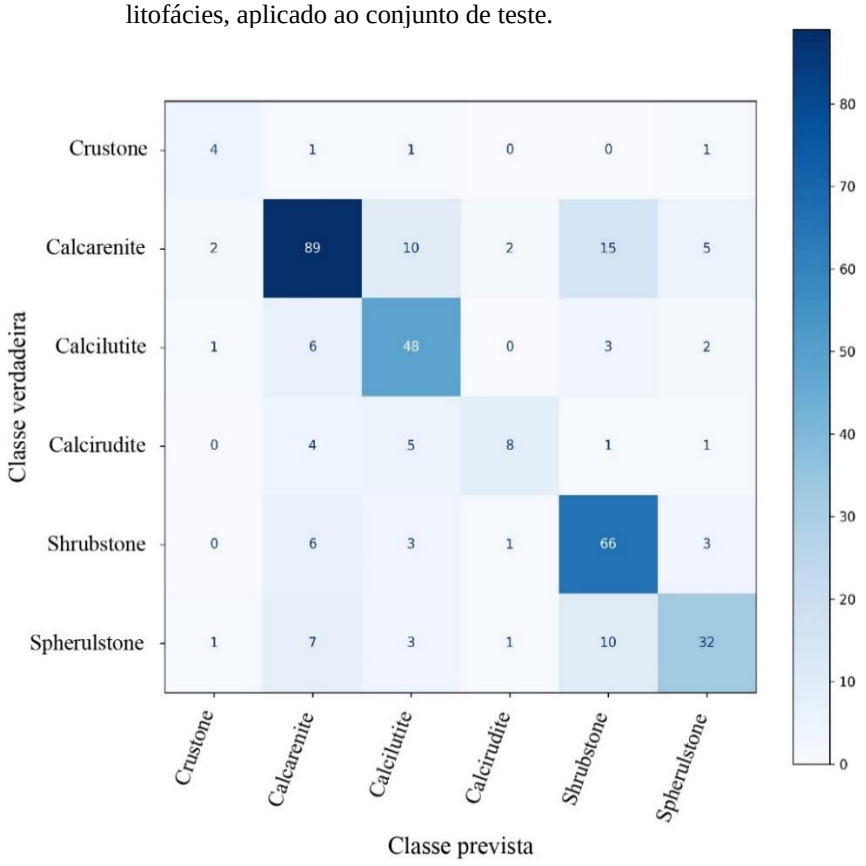
Fonte: A autora, 2023.

A análise da matriz de confusão revelou que o modelo de classificação 5 (Figura 13) apresentou diferentes taxas de erro ao confundir amostras de 9 classes de litofácies. A classe *spherulstone* foi incorretamente classificada como *calcarenite* e *shrubstone*. Observou-se que

uma amostra da classe *breccia* foi erroneamente prevista como *calcarenite*, e as amostras de *Chert* como *spherulstone* e *shrubstone*. Da mesma forma, a amostra de *dolostone* foi classificada como *spherulstone*. A classe *crustone* foi confundida com *calcarenite* e *spherulstone*. O erro predominante na classe *calcarenite* foi a classificação errônea como *shrubstone*, *spherulstone*, *calcirudite*, *calcilutite*, *crustone* e *breccia*. Erros na classe *calcilutite* incluíram classificações incorretas como *calcarenite*, *spherulstone*, *calcirudite* e *crustone*. As amostras de *calcirudite* foram erroneamente classificadas como *calcarenite*, *calcilutite*, *shrubstone* e *crustone*. A classe *shrubstone* apresentou a maior confusão associada à classe *calcarenite*, mas também relacionadas as classes *calcilutite*, *calcirudite* e *spherulstone*.

A matriz de confusão do modelo 9 (Figura 14) indicou que todas as classes foram preditas com uma certa taxa de acerto variável. As classes que tiveram melhor desempenho foram *calcarenite* com 78,76% e *spherulstone* com 72,73%. As classes com desempenho um pouco inferior, foram *calcilutite* com 68,57%, *shrubstone* com 69,47%, e *calcirudite* com 66,67%. A classe *crustone* teve o menor desempenho, com uma taxa de acerto de 50,00%.

Figura 14 - Matriz de confusão do modelo 9, composto por 6 classes de litofácies, aplicado ao conjunto de teste.



Fonte: A autora, 2023.

O exame da matriz de confusão do modelo 9 apontou diferentes taxas de erro ao confundir amostras das 6 classes de litofácies. Amostras da classe *crustone* foram erroneamente classificadas como *calcarenite, calcilutite* e *spherulstone*. Parte das mostras de *calcarenite* foram confundidas com todas as outras 5 classes, porém, com erro predominante associado às classes *calcilutite* e *shrubstone*. Erros na classe *calcilutite* incluíram classificações incorretas como *calcarenite, shrubstone, spherulstone* e *crustone*. Amostras de *calcirudite* foram erroneamente classificadas como *calcilutite, calcarenite, shrubstone* e *spherulstone*. A classe *shrubstone* apresentou a maior confusão associada à classe *calcarenite,* mas também relacionadas às classes *calcilutite*, *spherulstone* e *calcirudite*. Parte das mostras da classe *spherulstone* foram confundidas com todas as outras 5 classes, porém, com erro predominante associado às classes *shrubstone* e *calcarenite* (Figura 14).

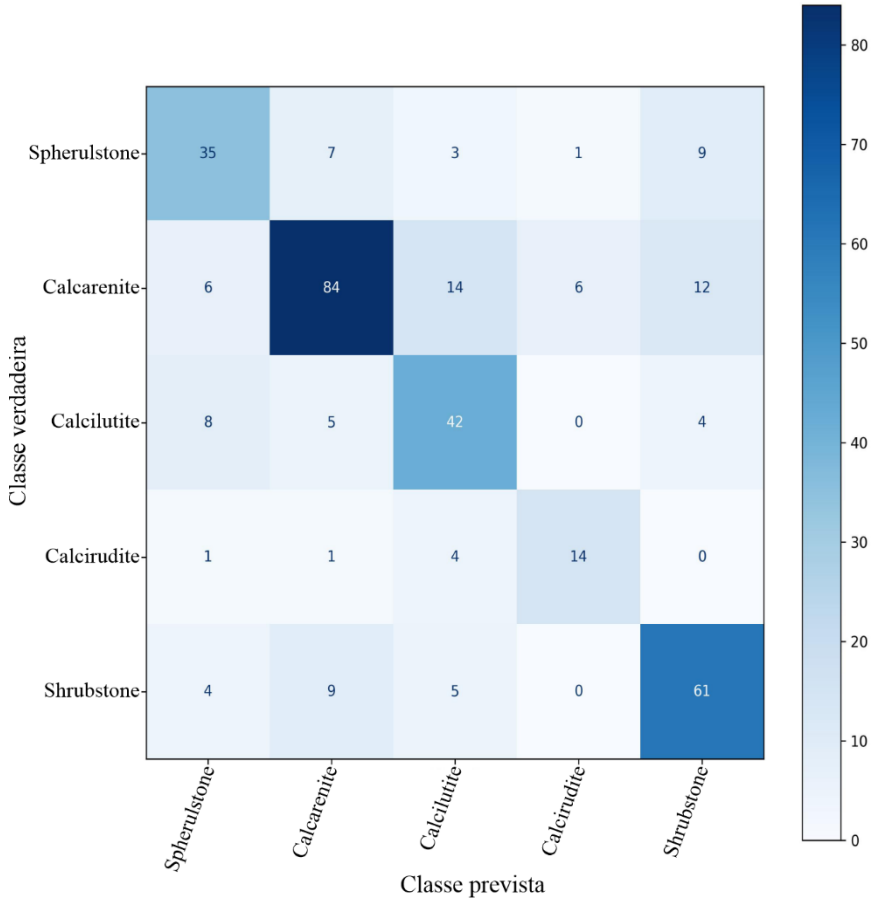
A matriz de confusão do modelo 11 (Figura 15) mostrou que todas as classes foram previstas corretamente. As classes que apresentaram o melhor desempenho foram *calcarenite*, com 79,25%, e *shrubstone*, com 70,93%. As classes com desempenho mais baixo, mas ainda satisfatório, foram *calcirudite* (66,67%), *spherulstone* (64,81%) e *calcilutite* (61,76%).

A avaliação da matriz de confusão do modelo 11 apontou diferentes taxas de erro ao confundir amostras das 5 classes de litofácies. Amostras da classe *spherulstone* foram confundidas com todas as outras 4 classes, com um erro predominante associado às classes *calcarenite* e *shrubstone*. Parte das amostras de *calcarenite* foram confundidas com todas as outras 4 classes, com um erro predominante associado às classes *calcilutite* e *shrubstone*. Erros na classe *calcilutite* incluíram classificações incorretas como *spherulstone, calcarenite* e *shrubstone*. Amostras de *calcirudite* foram erroneamente classificadas como *calcilutite*, *calcarenite* e *spherulstone*. A classe *shrubstone* apresentou a maior confusão associada à classe *calcarenite,* mas também foi confundida com as classes *calcilutite* e *spherulstone* (Figura 15).

A comparação entre os modelos revelou bom desempenho na identificação de diversas classes com número reduzido de amostras no conjunto de dados original, com exceção das classes *chert* e *dolostone*. Esta limitação pode indicar um problema com a representatividade no contexto da aplicação do SMOTE, como o número reduzido de amostras reais utilizadas como *k-neighbors*. Além disso, erros associados a algumas classes podem estar relacionados às similaridades nas características das litofácies da FBVE, que ocorrem frequentemente interdigitadas. Por exemplo, as amostras de *calcirudite-calcarenite* foram confundidas com *calcarenite,* e amostras de *calcilutite-calcarenite* com *calcarenite* e *calcilutite*. Nesses casos,

a falta de diversidade nos exemplos de treinamento pode ter resultado em uma generalização excessiva, onde o modelo não conseguiu distinguir nuances entre classes semelhantes.

Figura 15 - Matriz de confusão do modelo 11, composto por 5 classes de litofácies, aplicado ao conjunto e teste.



Fonte: A autora, 2023.

A classe majoritária no conjunto de dados, *calcarenite*, foi consistentemente bem classificada pelos modelos, destacando a importância da representatividade de classes no treinamento do algoritmo. Apesar da consistência na predição dessa classe, ela foi frequentemente confundida com as classes *shrubstone* e *calcilutite*, assim como a classe *shrubstone* com *calcarenite*. Confusões similares ocorreram com as classes *calcilutite*, *calcirudite*, *crustone* e *spherulstone*, também sugerindo que a possível ambiguidade nos rótulos devido à natureza dos dados geológicos, como a intercalação de alta frequência de litofácies e similaridades entre suas características, impactaram suas predições. Um conjunto de dados de treinamento não representativo das variações naturais das litofácies, ou ruídos nos

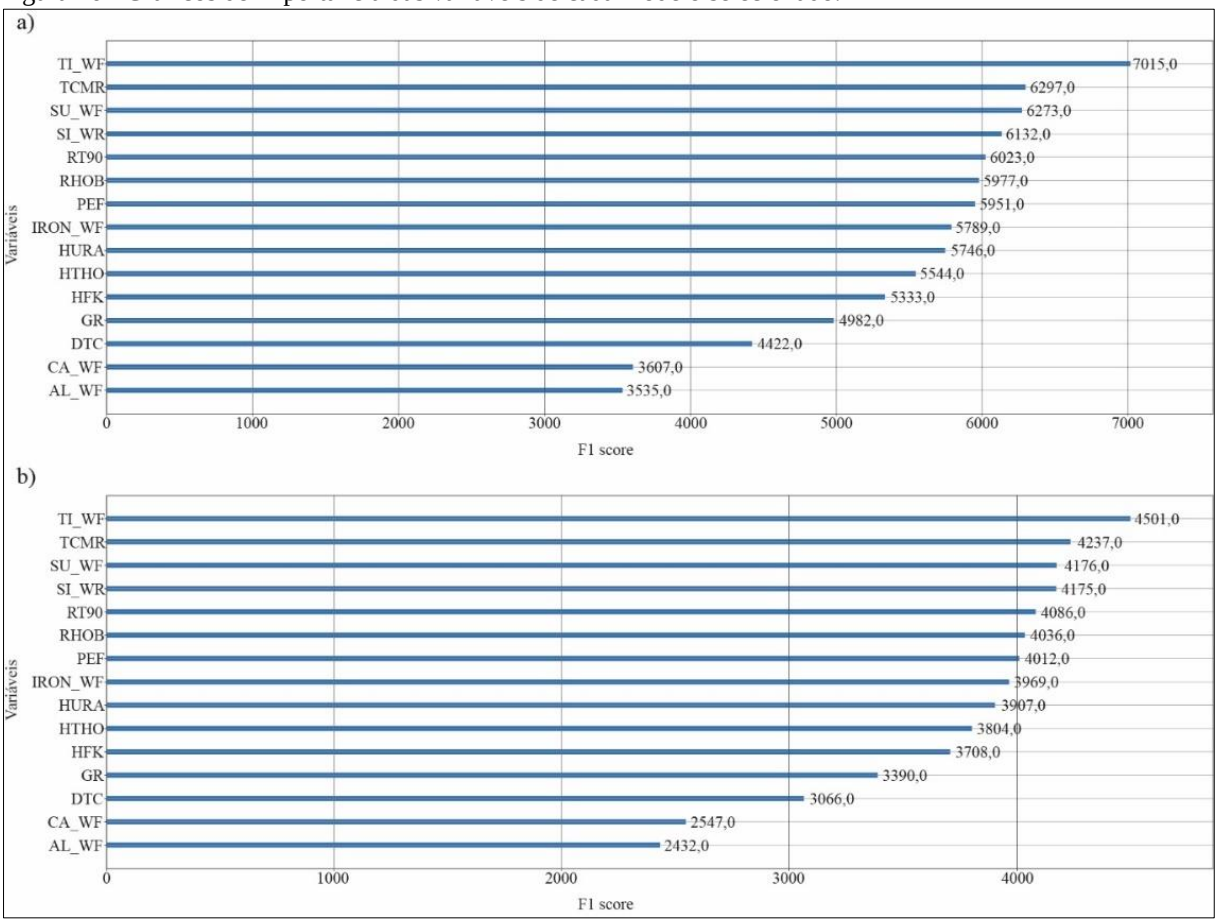
dados causados por erros ou tendenciosidades durante a sua aquisição, também podem ter contribuído para erros na classificação.

Cabe destacar que a semelhança entre as diversas classes de litofácies analisadas, de composição predominantemente carbonática, agregou complexidade ao problema proposto neste estudo, permitindo gerar modelos de classificação robustos no reconhecimento de pequenas diferenças entre os tipos de litofácies. Este aspecto pode ser comparado com os resultados obtidos por Gavidia et al. (2023), nos quais os autores se propuseram a classificar nove litofácies provenientes de poços que perfuraram a mesma formação. Ao analisar as classes no modelo desses autores, nota-se que eles parecem ter utilizado macroclasses de litofácies e litologias, simplificando, assim, o problema de classificação de litofácies. Grande parte dos estudos semelhantes também se propõem a classificar litologias (SUN et al., 2020) ou incluem apenas algumas litofácies focadas em um tipo específico de litologia (MEREMBAYEV et al., 2021), facilitando o processo.

Esta tese buscou avaliar a capacidade dos modelos de ML de reconhecer todas as litofácies encontradas (21 classes), ou subgrupos de litofácies semelhantes (11, 6 ou 5 classes). Além disso, ao contrário de alguns estudos analisados, nesta não foram inseridas características indicativas do contexto geológico em profundidade (HE; GU; XUE, 2022), nem o valor da profundidade real ou relativa ao topo da formação (SUN et al., 2020; MEREMBAYEV et al., 2021) como atributos de entrada para os modelos. Assim, entende-se que a inclusão futura de uma variável relacionada a essa posição pode melhorar os resultados, considerando que, dependendo da litofácies e de sua distribuição na formação, essa informação pode tornar os modelos mais robustos. No entanto, apesar da similaridade entre as litofácies, da ausência dessa variável e da quantidade reduzida de dados de algumas classes, os modelos apresentaram bom desempenho para a maioria das litofácies analisadas, o que é um resultado promissor para a continuação desta pesquisa.

A ordem de importância das variáveis utilizadas como entradas nos modelos selecionados é apresentada na Figura 16, que apresenta gráfico *feature importance* dos modelos 2 e 5. Essas variáveis apresentaram a mesma ordem de importância dos atributos que participam da estrutura das árvores de decisão dos diferentes modelos. Os perfis geofísicos mais utilizados foram o TI_WF (ECS - fração de peso de titânio), TCMR (NMR - porosidade total), SU_WF (ECS - fração de peso de enxofre) e SI_WF (ECS - fração de peso de silício). Em contraste, as variáveis menos empregadas incluíram GR (raios gama), DTC (*compressional slowness*), CA_WF (ECS - fração de peso de cálcio) e AL_WF (ECS - fração de peso de alumínio).

Figura 16 - Gráficos de importância das variáveis de cada modelo selecionado.



Legenda: (a) modelo 2; (b) modelo 5; TI_WF- ECS - fração de peso de titânio; TCMR- NMR - porosidade total; SU_WF- ECS - fração de peso de enxofre; SI_WF- ECS - fração de peso de silício; RT90- Resistividade profunda; RHOB- Densidade; PEF- Fator fotoelétrico; IRON_WF- ECS - fração de peso de ferro; HURA- GR espectral - concentração de urânio; HTHO- GR espectral - concentração de tório; HFK- GR espectral - concentração de potássio; GR- Raios gama; DTC- *Compressional slowness*; CA_WF- ECS - fração de peso de cálcio, e; AL_WF- ECS - fração de peso de alumínio.

Fonte: A autora, 2023.

Entre as variáveis mais influentes na construção dos modelos, três estão relacionadas à espectroscopia de captura elementar (ECS), uma ferramenta que mede o conteúdo elementar da rocha e auxilia na determinação da litologia e de outras propriedades das formações. O perfil TI_WF destacou-se significativamente na classificação. Segundo Alameedy (2023), concentrações elevadas desse perfil podem estar correlacionadas a dolomita ou intervalos dolomitizados em rochas carbonáticas. Já a variável SU_WF mostrou alta concentração na porção superior da FBVE, caracterizada pela intercalação de litofácies carbonáticas com evaporitos da Formação Ariri. Neste intervalo, a concentração de SU_WF aumenta de menos de 1% para mais de 20% em menos de 2 metros, em direção ao contato com a Formação Ariri. Essa formação é composta por litologias evaporíticas variadas, geralmente incluindo

halita e anidrita. Portanto, a anidrita, um sulfato de cálcio (CaSO4), pode ser a responsável pela alta concentração de SU_WF observada.

O perfil de SI-WF também se destacou na construção dos modelos. De Jesus et al. (2023) apontaram um aumento no volume de sílica da base ao topo da FBVE e associaram esse aumento à diminuição do volume de calcita. Este trabalho também identificou uma correlação negativa entre estas variáveis, que inclusive apresentaram contribuições opostas na classificação das litofácies. Embora todos os modelos testados tenham exibido a mesma ordem de relevância das variáveis, cada um apresentou variações únicas em suas pontuações de importância. Para averiguar o impacto dos atributos menos importantes, foi realizado um teste que consistiu na remoção destas variáveis e na repetição do processo de treinamento, validação e teste. Contudo, este experimento resultou em uma redução nas métricas de avaliação dos modelos, indicando que, apesar de serem consideradas características menos importantes, essas variáveis contribuíram positivamente para a construção dos modelos classificadores.

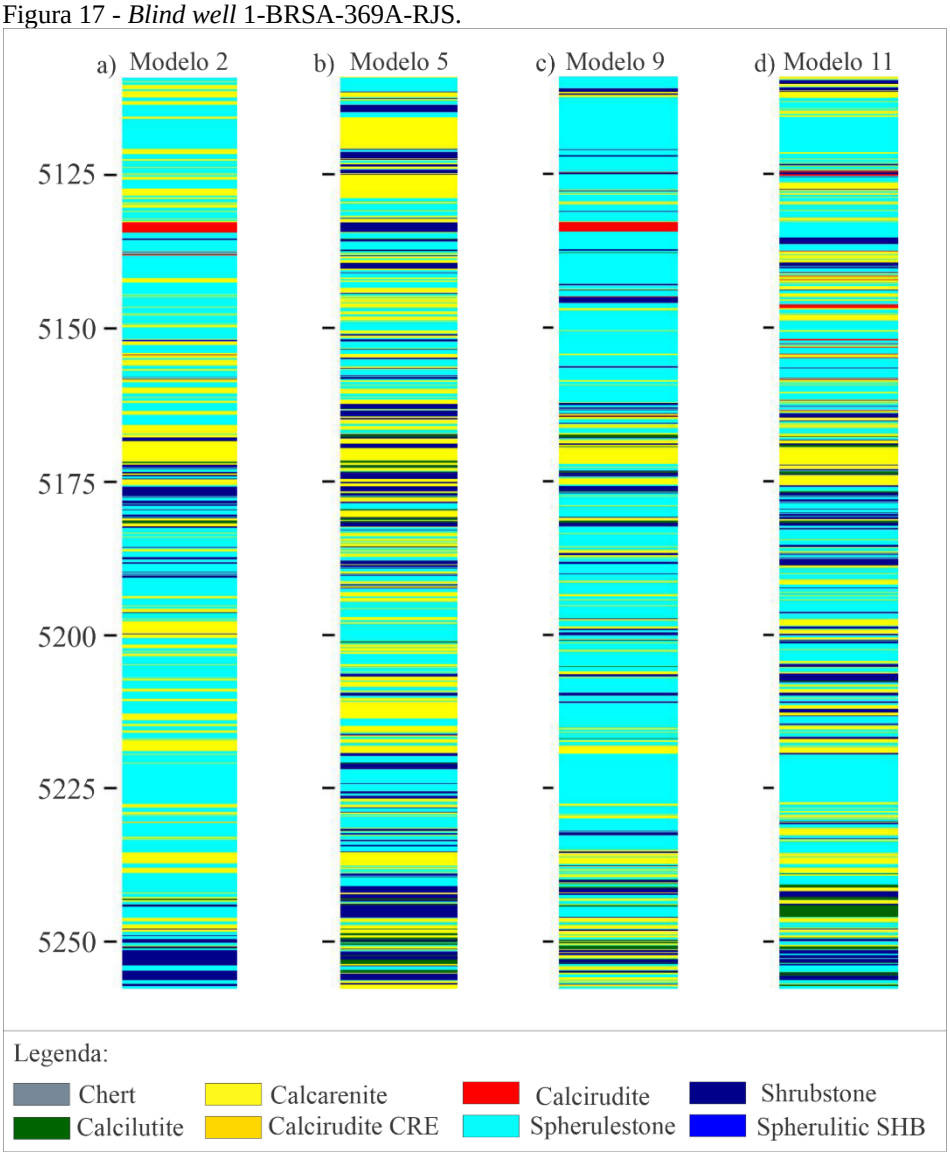
Por fim, como último resultado do presente estudo, os modelos de classificação de litofácies selecionados (modelos 2, 5, 9 e 11) foram aplicados na análise automatizada das litofácies de quatro *blind wells* (Tabela 1), que possuíam os mesmos dados geofísicos que os poços com rótulos de litofácies utilizados para treinamento, validação e teste dos modelos. Estes poços foram o 1-BRSA-369A-RJS, 3-BRSA-865A-RJS, 3-BRSA-883 e 3-BRSA-1120. Desta forma, mesmo que os resultados gerados não possam ser validados, pois não há informações sobre as litofácies encontradas nesses poços, estes são apresentados como um indicativo da aplicação do método proposto.

A análise dos resultados obtidos pela aplicação dos modelos classificadores no poço 1-BRSA-369A-RJS (Figura 17) revelou a possível presença de oito litofácies diferentes: *chert*, *calcarenite*, *calcirudite-calcarenite*, *calcilutite*, *calcirudite*, *shrubstone*, *spherulitic-shrubstone* e *spherulstone*. O modelo 2 conseguiu prever as oito litofácies mencionadas, o modelo 5 (Figura 17b) identificou quatro classes (*calcarenite*, *calcilutite*, *shrubstone* e *spherulstone)* enquanto os modelos 9 (Figura 17c) e 11 (Figura 17d) reconheceram cinco classes, com a inclusão do *calcirudite* que não foi classificado pelo modelo 5, de 11 classes.

A comparação entre os diferentes modelos aplicados ao poço 1-BRSA-369A-RJS apontou que todos foram consistentes na identificação das litofácies *calcarenite*, *shrubstone* e *spherulstone*, indicando uma concordância na interpretação geral das litofácies ao longo da profundidade do poço. Os modelos 2 e 9 foram consistentes ao identificar a litofácies

calcirudite no mesmo intervalo de profundidade, enquanto os modelos 5, 9 e 11 indicaram uma maior presença da litofácies *calcilutite* na porção mais basal do poço (Figura 17).

Figura 17 - *Blind well* 1-BRSA-369A-RJS.



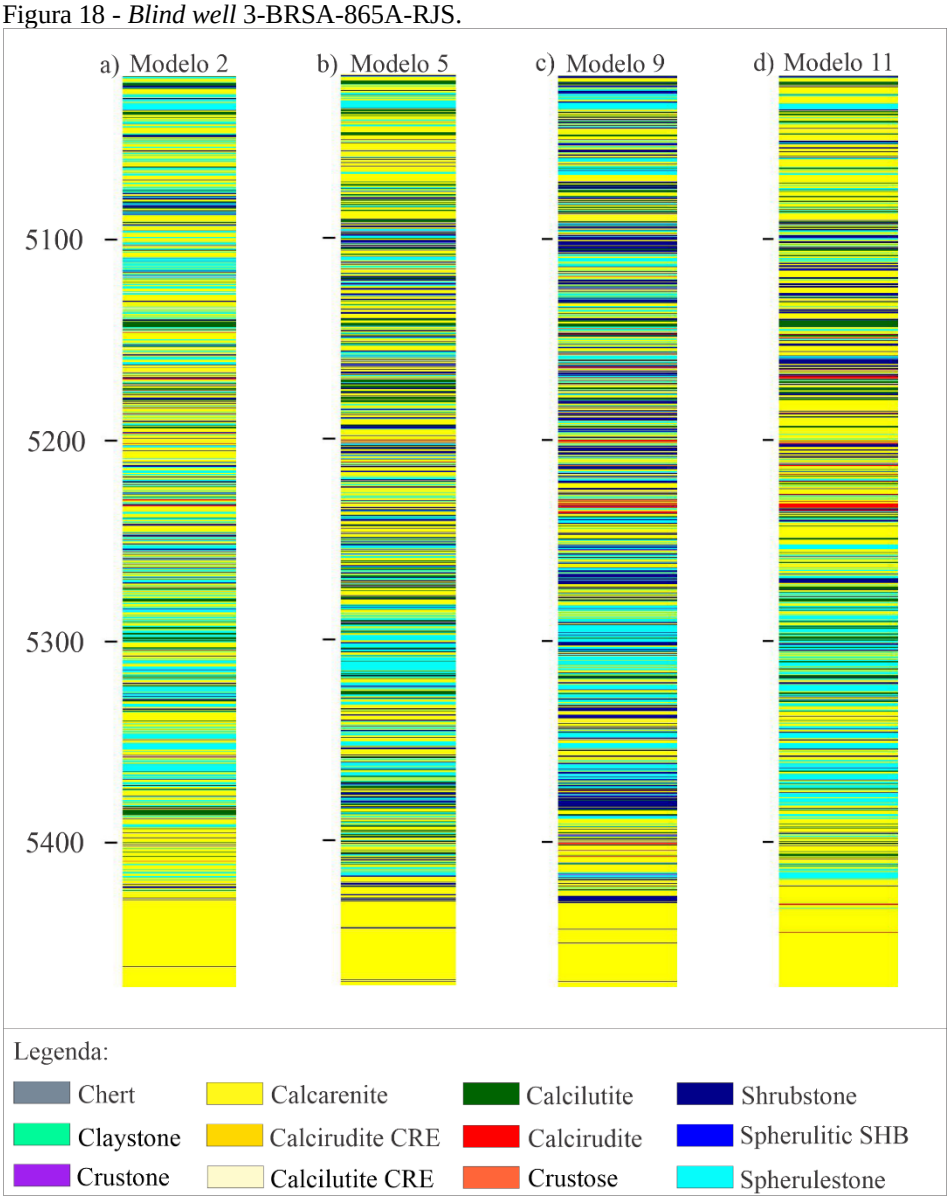
Legenda: CRE- *calcarenite*; SHB- *shrubstone*.
Fonte: A autora, 2023.

A aplicação dos modelos classificadores no poço 3-BRSA-865A-RJS (Figura 18a) indicou 12 litofácies no total: *chert*, *claystone*, *crustone*, *calcarenite*, *calcirudite-calcarenite*, *calcilutite-calcarenite*, *calcilutite*, *calcirudite*, *crustose*, *shrubstone*, *spherulitic-shrubstone* e *spherulstone*. O modelo 2 identificou onze litofácies, exceto *crustone*. O modelo 5 (Figura 18b) reconheceu seis classes: *chert*, *calcarenite*, *calcilutite*, *calcirudite*, *shrubstone* e *spherulstone*. O modelo 9 (Figura 18c) indicou seis classes: *crustone*, *calcarenite*, *calcilutite*,

calcirudite, *shrubstone* e *spherulstone*, e o modelo 11 (Figura 18d) cinco classes: *calcarenite*, *calcilutite*, *calcirudite*, *shrubstone* e *spherulstone*.

A análise das litofácies no poço 3-BRSA-865A-RJS (Figura 18) mostrou uma notável concordância entre todos os modelos na identificação das litofácies *calcarenite* e *spherulstone*, principalmente na região mais profunda do poço. Foi observada uma menor consistência na distribuição das litofácies *calcilutite* e *shrubstone*, indicando variações na interpretação dessas litofácies. Os modelos 9 e 11 identificaram consistentemente a litofácies *calcirudite* no mesmo intervalo de profundidade.

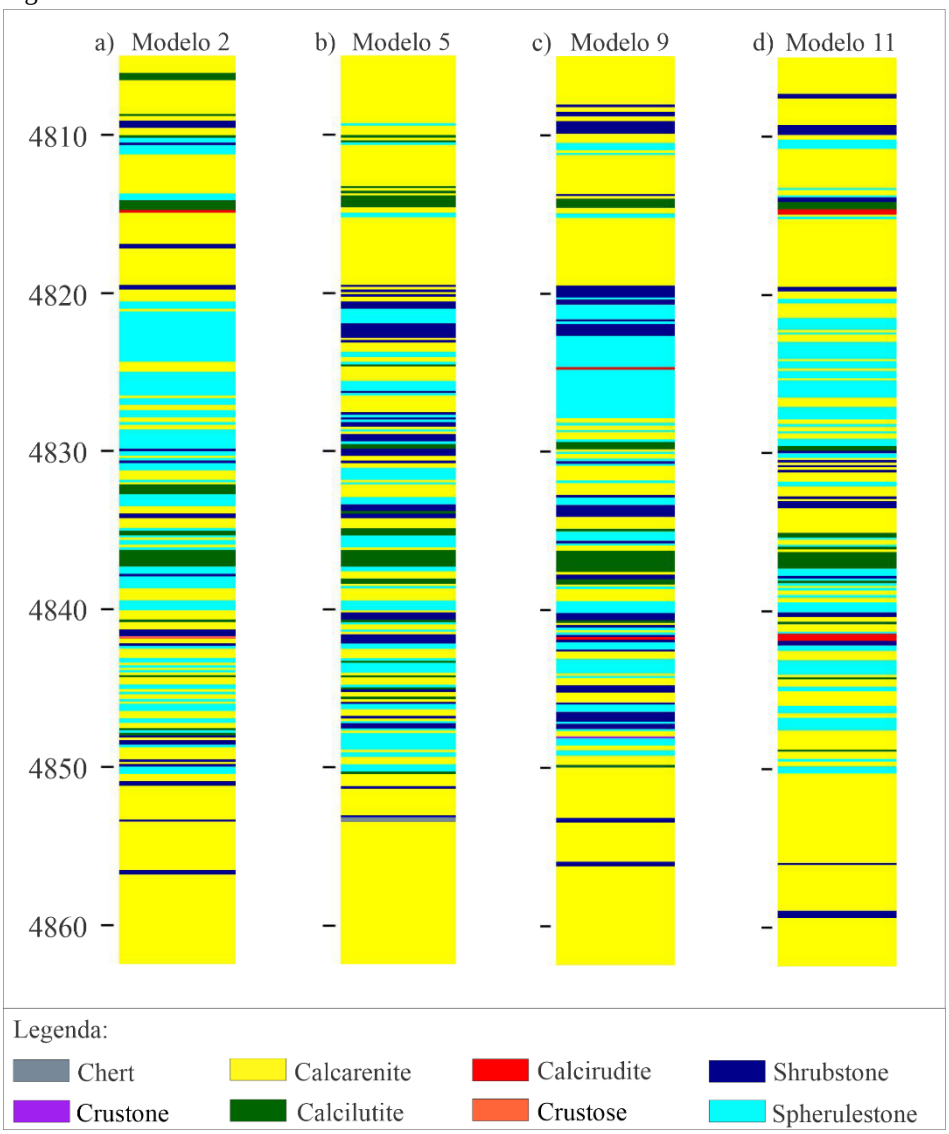
Figura 18 - *Blind well* 3-BRSA-865A-RJS.



Legenda: CRE- *calcarenite*; SHB- *shrubstone*; CRS- *crustone*.
Fonte: A autora, 2023.

No poço 3-BRSA-883-RJS, foram identificadas oito possíveis litofácies: *chert*, *crustone*, *calcarenite*, *calcilutite*, *calcirudite*, *crustose*, *shrubstone* e *spherulstone*. O modelo 2 (Figura 19a) identificou seis litofácies: *calcarenite*, *calcilutite*, *calcirudite*, *crustose*, *shrubstone* e *spherulstone*. O modelo 5 (Figura 19b) reconheceu cinco classes: *chert*, *calcarenite*, *calcilutite*, *shrubstone* e *spherulstone*. O modelo 9 (Figura 19c) indicou seis classes: *crustone*, *calcarenite*, *calcilutite*, *calcirudite*, *shrubstone* e *spherulstone*, e o modelo 11 (Figura 19d) cinco classes: *calcarenite*, *calcilutite*, *calcirudite*, *shrubstone* e *spherulstone*.

Figura 19 - *Blind well* 3-BRSA-883-RJS.



Legenda: CRS- *crustone*.
Fonte: A autora, 2023.

A avaliação das litofácies do poço 3-BRSA-883-RJS (Figura 19) revelou que todos os modelos identificaram uma predominância significativa da classe *calcarenite*, principalmente nas regiões de topo e da base do poço. Foi observada também uma consistência na identificação da litofácies *spherulstone* por todos os modelos entre as profundidades de 4.820 e 4.852 m. A litofácies *calcilutite* foi identificada por todos os modelos no intervalo de profundidade aproximado de 4.814 e 4.836 m. O Modelo 5 foi o que identificou o menor número de classes de litofácies. Os Modelos 2, 9 e 11 foram capazes de identificar a classe *calcirudite*, exibindo similaridades nas profundidades das suas previsões.

Por fim, a análise dos resultados no poço 3-BRSA-1120-RJS apontou a ocorrência de nove litofácies: *chert*, *crustone*, *calcarenite*, *calcilutite-calcarenite*, *calcilutite*, *calcirudite*, *dolostone*, *shrubstone* e *spherulstone*. O modelo 2 (Figura 20a) identificou oito litofácies: *chert*, *dolostone*, *calcarenite*, *calcilutite-calcarenite*, *calcilutite*, *calcirudite*, *shrubstone* e *spherulstone*. O modelo 5 (Figura 20b) reconheceu cinco classes: *calcarenite*, *calcilutite*, *calcirudite*, *shrubstone* e *spherulstone*. O modelo 9 (Figura 20c) indicou seis classes: *crustone*, *calcarenite*, *calcilutite*, *calcirudite*, *shrubstone* e *spherulstone*, e o modelo 11 (Figura 20d) cinco classes: *calcarenite*, *calcilutite*, *calcirudite*, *shrubstone* e *spherulstone*.

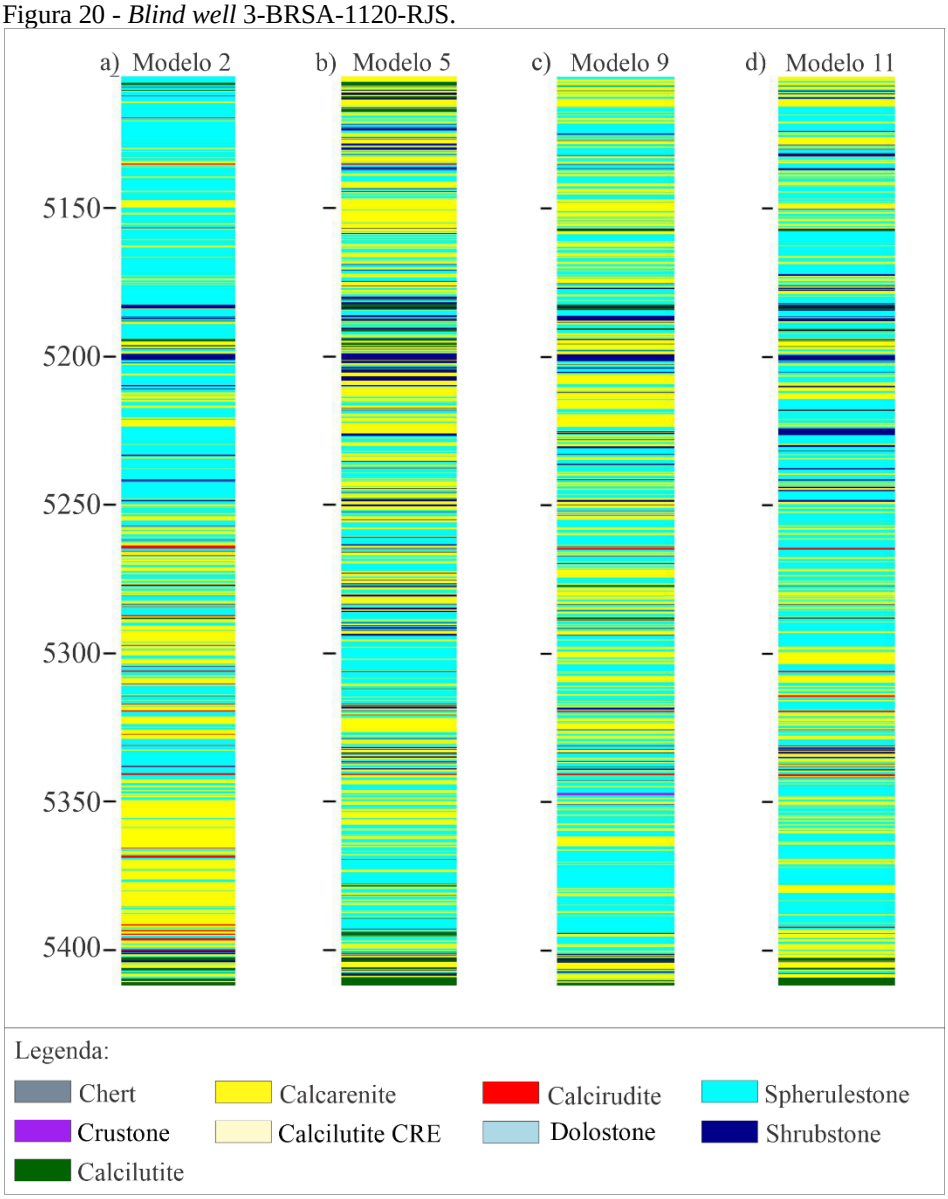
A análise das litofácies preditas do poço 3-BRSA-1120-RJS revelou a predominância das litofácies *spherulstone* e *calcarenite* em todos os modelos. Também foi observada uma consistência na identificação da litofácies *shrubstone* por todos os modelos entre as profundidades de 5.175 m e 5.205 m. Notou-se uma distribuição aleatória da litofácies *calcirudite*, principalmente abaixo dos 5.250 m. Na porção basal do poço, constatou-se a maior concentração da litofácies *calcilutite* a partir dos 5.400 m (Figura 20).

A análise indicou que, embora tenha havido concordância na identificação das classes de litofácies *calcarenite* e *spherulstone*, duas das principais classes utilizadas na construção dos modelos de classificação, houve variações notáveis na interpretação de outras litofácies principais. Essas classes incluem o *shrubstone*, *calcilutite* e *calcirudite*. Apesar das diferenças entre os modelos, foram identificados intervalos consistentes de litofácies, indicando a sensibilidade dos modelos na predição dessas classes. As variações entre eles podem ter sido devidas às diferenças nos modelos de classificação ou às diferenças implícitas nos dados geológicos e na gênese das litofácies identificadas na FBVE, que são fortemente interdigitadas.

A melhor performance do modelo 2 foi na sua aplicação ao poço 3-BRSA-865A-RJS, onde identificou onze classes. Seu pior desempenho foi no 3-BRSA-883-RJS, identificando 6 classes. O modelo 5 obteve seu melhor desempenho no poço 3-BRSA-865A-RJS onde

identificou 6 classes, enquanto seu pior desempenho ocorreu no 1-BRSA-369A-RJS, com a identificação de quatro classes. O modelo 9 identificou todas as seis classes em três poços, com exceção do 1-BRSA-369A-RJS. Finalmente, o modelo 11 demonstrou consistência ao apontar todas as cinco classes de litofácies em todos os *blind wells*.

Figura 20 - *Blind well* 3-BRSA-1120-RJS.



Legenda: CRE- *calcarenite*; CRS- *crustone*.
Fonte: A autora, 2023.

CONCLUSÕES

A análise realizada para desenvolver um fluxo de trabalho e aplicar uma técnica de *machine learning* supervisionada utilizando o algoritmo XGBoost demonstrou consistência na classificação de litofácies predominantemente carbonáticas da Formação Barra Velha no Campo de Tupi, Bacia de Santos. As etapas meticulosas de controle de qualidade e pré-processamento dos dados tiveram um impacto positivo nos resultados, destacando-se as estratégias de aumento artificial dos dados e o ajuste de hiperparâmetros através da busca aleatória (*Random Search*). Estas abordagens, em conjunto com outras detalhadas nesta tese, mostraram desempenho satisfatório na mitigação do impacto do desbalanceamento do conjunto de dados e contribuíram significativamente para o sucesso dos modelos.

Com a implementação dessas estratégias e a utilização de dados adequados, foi possível desenvolver quatro modelos de ML supervisionados utilizando o algoritmo XGBoost. Estes modelos de classificação automatizada apresentaram acurácias variando de 63 a 72%. Tais resultados são promissores e de grande relevância para os reservatórios do intervalo Pré-sal, conhecidos por sua heterogeneidade devido aos processos deposicionais e diagenéticos.

Os resultados alcançados demonstram o desenvolvimento de fluxos de trabalho eficientes para o tratamento, controle de qualidade e análise de grandes volumes de dados geológicos, visando uma abordagem mais precisa e integrada. Eles contribuem para reduzir significativamente o tempo necessário para a análise de dados geológicos complexos, oferecendo uma obtenção de dados rápida, precisa e com redução de viés. As estratégias aplicadas mostraram-se eficazes na superação dos desafios na caracterização das litofácies da Formação Barra Velha, auxiliando na classificação automatizada dessas litofácies únicas.

A semelhança e intercalação de alta frequência entre as diversas classes consideradas tornaram o problema proposto complexo e os modelos propostos robustos para distinguir mínimas diferenças entre as litofácies estudadas. O aumento na quantidade de dados/amostras disponíveis para treinamento tende a melhorar e refinar os modelos, indicando a necessidade de avaliação e inclusão de novos dados de treinamento para tornar os modelos mais robustos foi identificada como crucial.

É importante ressaltar que os modelos apresentados neste trabalho mostraram-se aplicáveis para a classificação automatizada de litofácies no Pré-sal da Bacia de Santos, Formação Barra Velha, Campo de Tupi. No entanto, o efeito que outros tipos de litofácies,

provenientes de outras regiões que não a estudada, teriam nos modelos treinados não foi avaliado. Portanto, os modelos gerados podem estar limitados a formações com características semelhantes às litofácies analisadas, como as da seção do Pré-sal das bacias de Campos e Kwanza, África. Para que o método proposto seja replicado para outras formações ou bacias sedimentares, provavelmente seria necessário que o conjunto de dados utilizados no treinamento dos modelos fosse representativo dessas novas áreas. Finalmente, esta tese apresentou uma nova abordagem capaz de aprimorar a classificação de litofácies da Formação Barra Velha no Campo de Tupi, essenciais para análises petrofísicas e sísmicas, contribuindo assim para a evolução do setor de energia.

Como trabalhos futuros, para aprimorar a pesquisa, recomendamos a ampliação do conjunto de dados de treinamento. Isso envolveria a inclusão de mais amostras de descrições de litofácies associadas a perfis geofísicos, tanto convencionais quanto avançados. Recomenda-se também a inclusão de variáveis que ofereçam uma compreensão mais profunda de contexto geológico. Outro aspecto importante é a comparação da variável TI_WF com as variáveis PEF e Mg_WF, com o objetivo de avaliar a sua eficácia na identificação de intervalos dolomitizados. Também é proposto um aprofundamento no estudo de hiperparâmetros para estabelecer uma lógica mais eficiente e aplicável às variações encontradas. Uma avaliação minuciosa da inclusão das litofácies crustone e calcitic-crustone em outras classes é igualmente crucial, visto que a abordagem que testou a remoção dessas classes impactou negativamente as métricas de avaliação dos modelos testados. Para melhorar os modelos de classificação, novas tentativas de otimização do algoritmo devem ser consideradas, bem como a introdução de outras técnicas ou variáveis. Por fim, sugere-se uma comparação entre o XGBoost e outras técnicas de Machine Learning supervisionado, com o objetivo de identificar uma abordagem mais eficiente, evitando fluxos de trabalho e dados de entrada excessivamente complexos.

REFERÊNCIAS

ABDOLAHI, A. et al. Seismic inversion as a reliable technique to anticipating of porosity and facies delineation, a case study on Asmari Formation in Hendijan field, southwest part of Iran. *Journal of Petroleum Exploration and Production Technology*, v. 12, n. 11, p. 3091–3104, 2022. Disponível em: <https://doi.org/10.1007/s13202-022-01497-y>.

ABELHA, M.; PETERSOHN, E. The State of the Art of the Brazilian Pre-Salt Exploration. *AAPG 2018 Annual Convention & Exhibition, Search and Discovery Article #30586 (2018)*, p. 1–43, 2018. Disponível em: <http://www.investidorpetrobras.com.br/download/1462>.

AL-MUDHAFA, W. J. Advanced supervised machine learning algorithms for efficient electrofacies classification of a carbonate reservoir in a giant southern iraqi oil field. *Proceedings of the Annual Offshore Technology Conference*, v. 2020- May, 2020.

ALAMEEDY, U. Accurate Petrophysical Interpretation of Carbonate using the Elemental Capture Spectroscopy (ECS). *Iraqi Journal of Chemical and Petroleum Engineering*, v. 24, n. 3, p. 125–131, 2023.

ALMEIDA, J. et al. Pre-rift tectonic scenario of the eo-cretaceous gondwana break-up along SE Brazil-SW Africa: Insights from tholeiitic mafic dyke swarms. *Geological Society Special Publication*, v. 369, n. 1, p. 11–40, 2013.

ALPAYDIN, E. *Introduction to machine learning*. Third Edit ed. MIT press, 2020.

ANP. *Boletim da Produção de Petróleo e Gás Natural ÍndiceBoletim da Produção de Petróleo e Gás Natural n° 158, 10/2023*. Disponível em: <https://www.gov.br/anp/pt-br/centrais-de-conteudo/publicacoes/boletins-anp/boletins/arquivos-bmppgn/2023/boletim-marco.pdf>.

BAHL, W. *Machine Learning: a Beginners Guide to History, Development and Future Possibilities of Machine Learning*. [s.l.] William Bahl, 2019.

BARNETT, A. J. et al. Origin and Significance of Thick Carbonate Grainstone Packages in Nonmarine Successions: A Case Study from the Barra Velha Formation, Santos Basin, Brazil. In: *AAPG Memoir 124: The Supergiant Lower Cretaceous Pre-Salt Petroleum Systems of the Santos*, p. 155–174, 2021.

BATISTA, G. E. A. P. A.; PRATI, R. C.; MONARD, M. C. A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD Explorations Newsletter*, v. 6, n. 1, p. 20–29, 2004.

BATISTA, G. E. A. P. A.; PRATI, R. C.; MONARD, M. C. Balancing strategies and class

overlapping. *Lecture Notes in Computer Science*, v. 3646, p. 24–35, 2005.

BERGSTRA, J.; BENGIO, Y. Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, v. 13, p. 281–305, 2012.

BORGHI, L. et al. Defining a new common language: a multi-scale descriptive classification for the pre-salt carbonates of the Barra Velha Formation. *Rio Oil & Gas Expo And Conference*, n. July, p. 1–10, 2022.

BRUCE, R. A Bayesian Approach to Semi-Supervised Learning. In: Nlprs2001, Asheville. *Anais*... Asheville: 2001. Disponível em: <http://www.afnlp.org/wp/>.

BUST, V. K.; OLETU, J. U.; WORTHINGTON, P. F. The challenges for carbonate petrophysics in petroleum resource estimation. *SPE Reservoir Evaluation and Engineering*, v. 14, n. 1, p. 25–34, 2011.

CARDOSO, P. H. A. *Fundamentos de Deep Learning na prática (parte 2)*. parte 2. 2020. Disponível em: < https://pedro-cardoso-36379.medium.com/fundamentos-de-deep-learning-na-pr%C3%A1tica-parte-2-eae9a2951d9f>.

CHANG, H. K. et al. Sistemas petrolíferos e modelos de acumulação de hidrocarbonetos na Bacia de Santos. *Revista Brasileira de Geociências*, v. 38, n. 2, p. 29–46, 2008.

CHAWLA, N. V et al. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, v. 16, p. 321–357, 2002. Disponível em: <https://dl.acm.org/doi/10.5555/1622407.1622416>.

CHEN, T.; GUESTRIN, C. XGBoost: A scalable tree boosting system. In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, *Anais*...2016.

CLENEAY, C. A. Log Analysis Applications. In: MORTON-THOMPSON, D.; WOODS, A. M. (Ed.). *Development Geology Reference Manual*. [s.l.] AAPG, 1992. p. 441–446.

COBBOLD, PETER ROBERT; MEISLING, KRISTIAN E; MOUNT, V. S. Segmentation of an obliquely-rifted margin, Campos and Santos basins, SE Brazil. *AAPG Bulletin*, v. 85, n. November, p. 1925–1944, 2001. Disponível em: <https://www.researchgate.net/publication/273337168_Segmentation_of_an_obliquely-rifted_margin_Campos_and_Santos_basins_SE_Brazil>.

COVER, T. M.; HART, P. E. Nearest Neighbor Pattern Classification. *IEEE Transactions on Information Theory*, v. 13, n. 1, p. 21–27, 1967.

DE JESUS, I. L. et al. Carbonate reservoir quality and permoporosity obliteration due to silicification processes in the Barra Velha Formation, Santos Basin, Southeastern Brazil. *Brazilian Journal of Geology*, v. 53, n. 2, p. 1–17, 2023.

DE ROS, L. F.; OLIVEIRA, D. M. An operational classification system for the South Atlantic pre-salt rocks. *Journal of Sedimentary Research*, p. 693–704, 2023.

DEV, V. A.; EDEN, M. R. Gradient Boosted Decision Trees for Lithology Classification. *Computer Aided Chemical Engineering*, v. 47, n. 2018, p. 113–118, 2019.

DHARMIK, R. C.; BAWANKAR, B. U. Design challenges for machine/deep learning algorithms. In: *Machine Learning Techniques for VLSI Chip Design*. [s.l.] John Wiley & Sons, Ltd, p. 195–209, 2023.

DOYEN, P. *Seismic Reservoir Characterization: An Earth Modelling Perspective (EET 2)Seismic Reservoir Characterization: An Earth Modelling Perspective (EET 2)*, 2007. . Disponível em: <https://www.eageseg.org/wp-content/uploads/2019/08/2007-EAGE-Education-Tour-EET-Seismic-Reservoir-Characterization-an-Earth-Modelling-Perspective.pdf>.

DRAHOKOUPIL, J. Application of the XGBoost algorithm and Bayesian optimization for the Bitcoin price prediction during the COVID-19 period. *FFA Working Paper*, v. 4, n. junho, 2022. Disponível em: <https://ideas.repec.org/p/prg/jnlwps/v4y2022id4.006.html>.

FACELI, K. et al. *Inteligeência Artificial: Uma Abordagem de Aprendizado de Máquina*. 2. ed. Rio de Janeiro: LTC - Livros Técnicos e Científicos Editora Ltda., 2023.

FALAVIGNA, G. P. et al. Controle de qualidade aplicado a dados gravimétricos. *Revista Brasileira de Geomática*, v. 2, n. 1, p. 20–29, 2014.

FARIAS, F. et al. Evaporitic carbonates in the pre-salt of Santos Basin – Genesis and tectonic implications. *Marine and Petroleum Geology*, v. 105, n. April, p. 251–272, 2019. Disponível em: <https://doi.org/10.1016/j.marpetgeo.2019.04.020>.

FONTANA, É. *Introdução aos Algoritmos de Aprendizagem Supervisionada*Curitiba, 2020. . Disponível em: <https://fontana.paginas.ufsc.br/>.

FONTELLES, M. J. et al. METODOLOGIA DA PESQUISA: DIRETRIZES PARA O CÁLCULO DO TAMANHO DA AMOSTRA. *Revista Paraense de Medicina*, v. 24, n. 2, p. 57–64, 2010.

FOOTE, K. D. *A Brief History of Machine Learning*. 2021. Disponível em: <https://www.dataversity.net/a-brief-history-of-machine-learning/>.

FRIEDMAN, J. Greedy Function Approximation : A Gradient Boosting Machine. *The Annals of Statistics*, v. 29, n. 5, p. 1189–1232, 2001. Disponível em: <https://www.jstor.org/stable/2699986>.

GARCÍA, V.; SÁNCHEZ, J. S.; MOLLINEDA, R. A. On the effectiveness of preprocessing methods when dealing with different levels of class imbalance. *Knowledge-Based Systems*, v.

25, n. 1, p. 13–21, 2012.

GAVIDIA, J. C. R. et al. Utilizing integrated artificial intelligence for characterizing mineralogy and facies in a pre-salt carbonate reservoir, Santos Basin, Brazil, using cores, wireline logs, and multi-mineral petrophysical evaluation. *Geoenergy Science and Engineering*, v. 231, n. PA, p. 212303, 2023. Disponível em: <https://doi.org/10.1016/j.geoen.2023.212303>.

GOÉS, C. B. D. *Aprendizado de máquina clássico e quântico: classificação de estados emaranhados e equações diferenciais parciais*. 2023. Universidade Federal de Santa Catarina, 2023.

GOMES, J. P. et al. Facies classification and patterns of lacustrine carbonate deposition of the Barra Velha Formation, Santos Basin, Brazilian Pre-salt. *Marine and Petroleum Geology*, v. 113, n. September 2019, p. 104176, 2020. Disponível em: <https://doi.org/10.1016/j.marpetgeo.2019.104176>.

GOMES, P. C. T. *Machine Learning para todos, de forma simples e com exemplos!*. 2019. Disponível em: <https://www.datageeks.com.br/machine-learning/>. Acesso em: 12 nov. 2023.

GU, Y. et al. Data-driven lithology prediction for tight sandstone reservoirs based on new ensemble learning of conventional logs: A demonstration of a Yanchang member, Ordos Basin. *Journal of Petroleum Science and Engineering*, v. 207, n. July, p. 109292, 2021. Disponível em: <https://doi.org/10.1016/j.petrol.2021.109292>.

GUEDES, T. A. et al. *Estatística DescritivaEstatística Descritiva*, 2005. .

HALOTEL, J.; DEMYANOV, V.; GARDINER, A. Value of Geologically Derived Features in Machine Learning Facies Classification. *Mathematical Geosciences*, v. 52, n. 1, p. 5–29, 2020. Disponível em: <https://doi.org/10.1007/s11004-019-09838-0>.

HE, H.; GARCIA, E. A. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, v. 21, n. 9, p. 1263–1284, 2009.

HE, M.; GU, H.; XUE, J. Log interpretation for lithofacies classification with a robust learning model using stacked generalization. *Journal of Petroleum Science and Engineering*, v. 214, n. August 2021, p. 110541, 2022. Disponível em: <https://doi.org/10.1016/j.petrol.2022.110541>.

IBRAHIM, B. et al. A novel XRF-based lithological classification in the Tarkwaian paleo placer formation using SMOTE-XGBoost. *Journal of Geochemical Exploration*, v. 245, n. May 2022, p. 107147, 2023. Disponível em: <https://doi.org/10.1016/j.gexplo.2022.107147>.

JAMAL, P. et al. 1-6 Data Normalization and Standardization: A Technical Report. *Machine*

Learning Technical Reports, v. 1, n. 1, p. 1–6, 2014. Disponível em: <https://docs.google.com/document/d/1x0A1nUz1WWtMCZb5oVzF0SVMY7a_58KQulqQVT8LaVA/edit#>.

KHANDELWAL, N. A brief introduction to XGBoost. *Towards Data Science*, p. 12, 2020. Disponível em: <https://towardsdatascience.com/a-brief-introduction-to-xgboost-3eaee2e3e5d6#:~:text=XGBoost vs Gradient Boosting,can be parallelized across clusters.>.

KIRA, K.; RENDELL, L. A. A Practical Approach to Feature Selection. In: Proceedings of the 9th International Workshop on Machine Learning, ICML 1992, *Anais*...1992.

KOSOLWATTANA, T. et al. A self-inspected adaptive SMOTE algorithm (SASMOTE) for highly imbalanced data classification in healthcare. *BioData Mining*, v. 16, n. 1, p. 1–14, 2023. Disponível em: <https://doi.org/10.1186/s13040-023-00330-4>.

KUMAR, T.; SEELAM, N. K.; RAO, G. S. Lithology prediction from well log data using machine learning techniques: A case study from Talcher coalfield, Eastern India. *Journal of Applied Geophysics*, v. 199, n. March, p. 104605, 2022. Disponível em: <https://doi.org/10.1016/j.jappgeo.2022.104605>.

LAVIER, L. L.; MANATSCHAL, G. A mechanism to thin the continental lithosphere at magma-poor margins. *Nature*, v. 440, n. 7082, p. 324–328, 2006.

LIKAS, A.; VLASSIS, N.; J. VERBEEK, J. The global k-means clustering algorithm. *Pattern Recognition*, v. 36, n. 2, p. 451–461, 2003.

LISTER, G. S.; DAVIS, G. A. The origin of metamorphic core complexes and detachment faults formed during Tertiary continental extension in the northern Colorado River region, U.S.A. *Journal of Structural Geology*, v. 11, n. 1–2, p. 65–94, 1989. Disponível em: <https://www.sciencedirect.com/science/article/pii/0191814189900369>.

LISTER, G. S.; ETHERIDGE, M. A.; SYMONDS, P. A. Detachment faulting and the evolution of passive continental margins. *Geology*, v. 14, n. 3, p. 246–250, 1986.

LUCIA, F. J.; KERANS, C.; JENNINGS, J. W. Technology Today Series Carbonate Reservoir Characterization. *Society of Petroleum Engineers - SPE Asia Pacific Oil and Gas Conference and Exhibition 2010, APOGCE 2010*, n. June, p. 70–72, 2003.

MALKI, M. L. et al. Underlying mechanisms and controlling factors of carbonate reservoir characterization from rock physics perspective: A comprehensive review. *Geoenergy Science and Engineering*, v. 226, 2023. Disponível em: <https://doi.org/10.1016/j.geoen.2023.211793>.

MATIAS, H. C. et al. Unlocking Pandora - Insights from pre-salt reservoirs in Campos and Santos Basins (offshore Brazil). In: 77th EAGE Conference and Exhibition 2015: Earth

Science for Energy and Environment, July, Madrid. *Anais*... Madrid: 2015.

MCKENZIE, D. Some remarks on the development of sedimentary basins. *Earth and Planetary Science Letters*, v. 40, n. 1, p. 25–32, 1978.

MEREMBAYEV, T. et al. A Comparison of Machine Learning Algorithms in Predicting Lithofacies: Case Studies from Norway and Kazakhstan. *Energies*, v. 14, n. 7, p. 1–16, 2021.

MEREMBAYEV, T.; YUNUSSOV, R.; YEDILKHAN, A. Machine learning algorithms for classification geology data from well logging. In: 14th International Conference on Electronics Computer and Computation, ICECCO 2018, *Anais*...IEEE, 2018.

MILANI, E. J.; THOMAZ FILHO, A. Sedimentary Basins of the South America. *Tectonic Evolution of South America*, n. November, p. 389–449, 2000.

MIOT, H. A. Análise de sobrevivência em estudos clínicos e experimentais. *Jornal Vascular Brasileiro*, v. 16, n. 4, p. 267–269, 2017.

MME; EPE. Balanço Energético Nacional Relatório Síntese 2023. *Relatório de sintese*, p. 65, 2023.

MOHRIAK, W. U. et al. Crustal architecture of South Atlantic volcanic margins. *Special Paper of the Geological Society of America*, v. 362, p. 159–202, 2002.

MOHRIAK, W. U. Bacias da Margem Continental Divergente. In: HASUI, Y. et al. (Ed.). *Geologia do Brasil*. Primeira e ed. São Paulo: Editora Beca, 2012a. p. 466–480, 2012.

MOHRIAK, W. U. Bacias de Santos, Campos e Espírito Santo. In: HASUI, Y. et al. (Ed.). *Geologia do Brasil*. Primeira e ed. São Paulo: Editora Beca, 2012b. p. 481–496, 2012.

MORAVEC, H. P. *Obstacle Avoidance and Navigation in the Real World by a Seeing Robot Rover, Tech. Report, Robotics Institute, Carnegie-Mellon University, pp.1-175*. 1980. 1980. Disponível em: <http://oai.dtic.mil/oai/oai?verb=getRecord&metadataPrefix=html&identifier=ADA092604>.

MOREIRA, J. et al. Bacia de Santos. *Boletim de Geociências da Petrobras*, v. 15, 2007.

MOULIN, M.; ASLANIAN, D.; UNTERNEHR, P. A new starting point for the South and Equatorial Atlantic Ocean. *Earth-Science Reviews*, v. 98, n. 1–2, p. 1–37, 2010. Disponível em: <http://dx.doi.org/10.1016/j.earscirev.2009.08.001>.

MURPHY, K. P. *Machine learning: a probabilistic perspective*. London: The MIT Press, 2012.

NEVES, I. D. A. et al. Presalt reservoirs of the Santos Basin: Cyclicity, electrofacies, and tectonic-sedimentary evolution. *Interpretation*, v. 7, n. 4, p. SH33–SH43, 2019.

NICHOLS, G. *Livros*. 2nd. ed. [s.l.] John Wiley & Sons, Ltd, 2009.

PARSA, M. A data augmentation approach to XGboost-based mineral potential mapping: An

example of carbonate-hosted Zn–Pb mineral systems of Western Iran. *Journal of Geochemical Exploration*, v. 228, n. March, p. 106811, 2021. Disponível em: <https://doi.org/10.1016/j.gexplo.2021.106811>.

PEQUENO, H. C. *Classificação De Eletrofácies Em Estágios Turbidíticos De 3ª E 4ª Ordens Do Membro Caruaçu Da Formação Maracangalha, Um Estudo De Caso Do Campo De Massapê, Bacia Do Recôncavo*. 2019. Universidade Federal Fluminense, 2019.

PEREIRA, T. P. et al. Distribution of silicification intervals throughout the Barra Velha and Itapema formations: Host rock controls and chronology of silica precipitation (Pre-Salt, Santos Basin, Brazil). *Journal of South American Earth Sciences*, v. 128, n. March, 2023.

PRATI, R. C.; BATISTA, G. E. A. P. A.; MONARD, M. C. Class imbalances versus class overlapping: An analysis of a learning system behavior. *Lecture Notes in Artificial Intelligence*, v. 2972, p. 312–321, 2004.

RAGB, H. et al. Convolutional neural network based on transfer learning for breast cancer screening. n. January, 2021. Disponível em: <http://arxiv.org/abs/2112.11629>.

RAMADHAN, A. A. et al. Evaluation of Petrophysical Properties Using Imaging Techniques. *IOP Conference Series: Materials Science and Engineering*, v. 579, n. 1, 2019.

RIDER, M. H. *Geological Interpretation of Well Logs*, 2nd. ed. [s.l.] Rider French Consulting, Ltd. 2002.

RIGOTI, C. A. *Universidade do Estado do Rio de Janeiro Centro de Tecnologia e Ciências Faculdade de Geologia Caesar Augusto Rigoti Evolução tectônica da Bacia de Santos com ênfase na geometria crustal : Interpretação integrada de dados de sísmica de reflexão e refração*. 2015. Universidade do Estado do Rio de Janeiro, 2015. Disponível em: <https://www.bdtd.uerj.br:8443/handle/1/7131>.

SALLER, A. et al. Presalt stratigraphy and depositional systems in the Kwanza Basin, offshore Angola. *AAPG Bulletin*, v. 100, n. 7, p. 1135–1164, 2016.

SAMUEL, A. L. Some Studies in Machine Learning. *IBM Journal of Research and Development*, v. 3, n. 3, p. 210–229, 1959. Disponível em: <https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5392560>.

SANCHES, M. K. *Aprendizado de máquina semi-supervisionado: proposta de um algoritmo para rotular exemplos a partir de poucos exemplos rotulados*. 2003. USP, 2003. Disponível em: <https://teses.usp.br/>.

SHANNON, C. E. A Mathematical Theory of Communication. *Bell System Technical Journal*, v. 27, n. 4, p. 379–423, 623–656, 1948.

SILVA, S. F. C. R. da et al. Evolução Tectonoestratigráfica Da Formação Barra Velha Na

Área Dos Campos De Lapa E Sapinhoá, Bacia De Santos-Brasil Tectonostratigraphic Evolution of Barra Velha Formation in the Lapa and Sapinhoá Oil Fields, Santos Basin-Brazil. *Revista Geociências - UNESP*, v. 40, n. 1, p. 55–69, 2021. Disponível em: <https://www.periodicos.rc.biblioteca.unesp.br/index.php/geociencias/article/view/14679>.

STOLLHOFEN, H. et al. Tectonic and volcanic controls on Early Jurassic rift-valley lake deposition during emplacement of Karoo flood basalts, southern Namibia. *Palaeogeography, Palaeoclimatology, Palaeoecology*, v. 140, n. 1–4, p. 185–215, 1998.

SUN, Z. et al. A data-driven approach for lithology identification based on parameter-optimized ensemble learning. *Energies*, v. 13, n. 15, p. 1–15, 2020.

TERRA, Á. de O.; FERREIRA, A. da S.; OLIVEIRA, D. C. Os desafios do pré-sal brasileiro: um estudo da logística do campo de Tupi. In: Anais XI Simposio de Excelência em gestão e tecnologia, Rio de Janeiro. *Anais*... Rio de Janeiro: 2014. Disponível em: <https://www.aedb.br/seget/arquivos/artigos14/47420570.pdf>.

TERRA, J. G. S. et al. Classificações Clássicas De Rochas Carbonáticas. In: *Boletim de Geociências Petrobras*, 18p. 9–29, 2010.

VAN DER PLUIJM, BEN; MARSHAK, S. *Earth Structure: An Introduction to Structural Geology and Tectonics*. [s.l.] WW Norton&Company.Inc. pp, p. 270-277, 2004.

VISA, S. et al. Edited by Sofia Visa, Atsushi Inoue, and Anca Ralescu. *Maics*, v. 710, p. 120–127, 2011.

WERNICKE, B. Uniform-sense normal simple shear of the continental lithosphere. *Canadian Journal of Earth Sciences*, v. 22, n. 1, p. 108–125, 1985.

WRIGHT, V. P.; BARNETT, A. J. An abiotic model for the development of textures in some South Atlantic early Cretaceous lacustrine carbonates. *Geological Society Special Publication*, v. 418, n. 1, p. 209–219, 2015.

WRITHT, P. V; BARNETT, A. J. Critically Evaluating the Current Depositional Models for the Pre-Salt Barra Velha Formation, Offshore Brazil. *Search and Discovery*, v. 418, n. 1, 2017.

WU, F. X. Genetic weighted k-means algorithm for clustering large-scale gene expression data. *BMC Bioinformatics*, v. 9, n. SUPPL. 6, 2008.

XGBOOST. *XGBoost Parameters*. 2023. Disponível em: <https://xgboost.readthedocs.io/en/stable/parameter.html>. Acesso em: 12 dez. 2023.

ZALÁN, P. V. et al. An entirely new 3-D view of the crustal and mantle structure of a South Atlantic Passive Margin, Santos, Campos and Espírito Santo Basins, Brazil. In: AAPG Annual Conference and Exhibition, *Anais*...2011.

ZHANG, L.; ZHAN, C. Machine Learning in Rock Facies Classification: An Application of XGBoost. p. 1371–1374, 2017.