UNIVERSIDADE ESTADUAL DE CAMPINAS
Instituto de Geociências

HENRIQUE MOREIRA SANTANA

MODELAGEM GEOLÓGICA 3D DE UM RESERVATÓRIO CARBONÁTICO DA BACIA DE SANTOS UTILIZANDO APRENDIZAGEM DE MÁQUINA

CAMPINAS
2024

HENRIQUE MOREIRA SANTANA

MODELAGEM GEOLÓGICA 3D DE UM RESERVATÓRIO CARBONÁTICO DA BACIA DE SANTOS UTILIZANDO APRENDIZAGEM DE MÁQUINA

DISSERTAÇÃO APRESENTADA AO INSTITUTO DE GEOCIÊNCIAS DA UNIVERSIDADE ESTADUAL DE CAMPINAS PARA OBTENÇÃO DO TÍTULO DE MESTRE EM GEOCIÊNCIAS NA ÁREA DE GEOLOGIA E RECURSOS NATURAIS

ORIENTADOR: PROF. DR. EMILSON PEREIRA LEITE

ESTE EXEMPLAR CORRESPONDE À VERSÃO FINAL DA DISSERTAÇÃO DEFENDIDA PELO ALUNO HENRIQUE MOREIRA SANTANA E ORIENTADA PELO PROF. DR. EMILSON PEREIRA LEITE

CAMPINAS

2024

Ficha catalográfica
Universidade Estadual de Campinas (UNICAMP)
Biblioteca do Instituto de Geociências
Marta dos Santos - CRB 8/5892

Santana, Henrique Moreira, 1996-
Sa59m Modelagem geológica 3D de um reservatório carbonático da Bacia de Santos utilizando aprendizagem de máquina / Henrique Moreira Santana. – Campinas, SP : [s.n.], 2024.

Orientador: Emilson Pereira Leite.
Dissertação (mestrado) – Universidade Estadual de Campinas (UNICAMP), Instituto de Geociências.

1. Reservatórios de petróleo. 2. Pré-sal. 3. Inteligência artificial. 4. Modelagem tridimensional. I. Leite, Emilson Pereira, 1975-. II. Universidade Estadual de Campinas (UNICAMP). Instituto de Geociências. III. Título.

Informações Complementares

Título em outro idioma: 3D geological modeling of a carbonate reservoir in the Santos Basin using machine learning
Palavras-chave em inglês:
Petroleum reservoirs
Pre-salt
Artificial intelligence
Three-dimensional modeling
Área de concentração: Geologia e Recursos Naturais
Titulação: Mestre em Geociências
Banca examinadora:
Emilson Pereira Leite [Orientador]
Alexandre Campane Vidal
Wagner Moreira Lupinacci
Data de defesa: 30-04-2024
Programa de Pós-Graduação: Geociências

Identificação e informações acadêmicas do(a) aluno(a)
- ORCID do autor: https://orcid.org/0009-0000-6652-6820
- Currículo Lattes do autor: http://lattes.cnpq.br/7162362649997720

UNIVERSIDADE ESTADUAL DE CAMPINAS

INSTITUTO DE GEOCIÊNCIAS

AUTOR: Henrique Moreira Santana

MODELAGEM GEOLÓGICA 3D DE UM RESERVATÓRIO CARBONÁTICO DA BACIA DE SANTOS UTILIZANDO APRENDIZAGEM DE MÁQUINA

ORIENTADOR: Prof. Dr. Emilson Pereira Leite

Aprovado em: 30 / 04 / 2024

EXAMINADORES:

Prof. Dr. Emilson Pereira Leite - Presidente

Prof. Dr. Alexandre Campane Vidal

Prof. Dr. Wagner Moreira Lupinacci

A Ata de Defesa assinada pelos membros da Comissão Examinadora consta no processo de vida acadêmica do aluno.

Campinas, 30 de Abril de 2024.

SÚMULA/BIOGRAFIA

Geólogo pela Unicamp com experiências acadêmicas na área de geotecnologias aplicadas ao estudo dos recursos naturais. Participei de projetos de iniciação científica que envolveram: (i) a modelagem geológica 3D de um depósito mineral de Zn/Pb localizado no Vale do Ribeira; (ii) um estudo de aplicação de geotermometria de clorita na exploração de Au em um depósito da Província Mineral de Alta Floresta; (iii) a integração de dados de suscetibilidade magnética da cratera de impacto de Araguainha; e (iv) a aquisição de dados de gamaespectrometria e suscetibilidade magnética de amostras de coleções didáticas do IG/Unicamp.

Fui estagiário do Laboratório de Resíduos e Áreas Contaminadas do IPT (Instituto de Pesquisas Tecnológicas do Estado de São Paulo), onde também atuei como Pesquisador Visitante. No IPT, tive experiências com projetos de investigação e confirmação de áreas com suspeita de contaminação, realização de ensaios laboratoriais de determinação de parâmetros físicos do solo (e.g. granulometria, teor de umidade, condutividade hidráulica etc.) e com transformação digital aplicada às geociências.

Durante o mestrado, tive experiências de Estágio Docente nas disciplinas de Geofísica e de História das Ciência Naturais, oferecidas pelo Instituto de Geociências da Unicamp para alunos de graduação.

Meu mestrado em Geociências na Unicamp envolveu o estudo da aplicação de aprendizagem de máquina na modelagem geológica de reservatórios de hidrocarbonetos. A área de estudo consistiu no Campo de Tupi, localizado na Bacia de Santos. Esta pesquisa de mestrado fez parte do projeto de P&D entre UNICAMP/Shell Brasil/ANP (CW266675) denominado "Análise integrada multi-escala de rochas carbonatadas do pré-sal para a caracterização e previsão das propriedades dos reservatórios" e contou com bolsa acadêmica financiada pela Shell Brasil através da Fundação de Desenvolvimento da UNICAMP (FUNCAMP). Parte dos resultados dessa pesquisa foram apresentados no 84º EAGE Annual Conference & Exhibition.

AGRADECIMENTO

O presente trabalho foi realizado com apoio da Shell Brasil por meio do projeto de P&D registrado ANP nº 21575-6, intitulado “Análise Integrada de Multi-Escala de Rochas Carbonáticas do Pré-Sal para Caracterização e Predição de Propriedades de Reservatórios” (Unicamp/Shell Brasil/ANP), patrocinado pela Shell Brasil no âmbito da taxa de P&D da ANP como “Compromisso de Investimentos com Pesquisa e Desenvolvimento”.

Agradeço à Geosoftware por fornecer a licença do software Jason Workbench, à ANP pelo conjunto de dados de perfil do poço, e à Shell Brasil pelo conjunto de dados sísmicos.

Em especial, quero agradecer:

Aos meus pais, Lucivânia e Mauro, e à toda minha família pelo apoio incondicional. Nunca termino de aprender com vocês!

Ao meu orientador, Emilson, por todo conhecimento compartilhado, pela paciência, pela ajuda nos momentos difíceis, e pelo exemplo de *como orientar*!

Aos colegas do Projeto Shell: Jadson, Bruna, Daniel, Marta, Nathalia, Taynah e Rafael. A pesquisa pode ser, em muitos momentos, uma atividade solitária. Mas tive a sorte de poder contar com a ajuda de vocês neste trabalho. Muito obrigado!

À Janaina Büll por todo apoio, acolhimento e ensinamentos que me permitiram não apenas concluir este trabalho, mas também a enxergar a vida de outra forma.

Aos meus amigos da República Viracopos *(minha segunda casa em Campinas)*, da geologia, de Teixeira de Freitas e, em especial, a João Calil, Leonardo Reis e Gabriel Castanheira por todos os momentos compartilhados. A amizade de vocês foi fundamental para o andamento desta pesquisa.

Aos trabalhadores do IG-Unicamp, especialmente os professores Gelvam, Emilson e Jefferson pelas oportunidades de estágio docente (PED) nas disciplinas de Geofísica e História das Ciências Naturais. Jeff, obrigado por toda experiência compartilhada através do projeto de extensão “Ciclo das Águas no Marielle Vive”.

Aos camaradas do Coletivo Dínamo de Engenharia Popular por todo aprendizado através dos projetos de extensão popular.

Por fim, agradeço ao coletivo Capoeira Para Todes, ao Forró da Casa do Lago e à Bateria Alcalina pelas experiências que tanto acrescentaram em minha vida e foram fundamentais para a minha saúde durante esta pesquisa.

RESUMO

A modelagem geológica tridimensional de um campo petrolífero é fundamental para o planejamento das etapas de exploração e produção de hidrocarbonetos. Alguns produtos comuns da modelagem de reservatórios são os modelos 3D de fácies petroelásticas, unidades de fluxo, unidades geomecânicas e tipos de poros. Comumente, o entendimento da distribuição dessas classes em subsuperfície é adquirido por meio da integração de dados obtidos por levantamentos sísmicos e pela perfilagem geofísica de poços. Entretanto, apesar dos avanços na determinação dessas classes com certa precisão para cada poço, há o problema da classificação no espaço entre os poços. Em geral, essa classificação é feita por meio da marcação arbitrária de zonas homólogas em diagramas de dispersão dos parâmetros elásticos, como razão de velocidades das ondas P e S (Vp/Vs) e impedância acústica da onda P (Zp). Contudo, a região interpretada no volume sísmico depende criticamente do tamanho da zona selecionada no diagrama de dispersão, tornando o processo muito subjetivo. Posto isto, a abordagem seguida neste estudo consiste na classificação de unidades a partir de um algoritmo baseado em aprendizagem de máquina. Dessa forma, é eliminada a necessidade de se definir arbitrariamente um limiar para a separação de zonas homólogas. O algoritmo de classificação supervisionada foi treinado a partir de parte dos poços, utilizando 1 poço para teste até que todos tenham sido testados. O estudo avaliou a melhoria de desempenho do modelo ao incorporar aos dados de treino pseudocurvas extraídas dos produtos de inversão sísmica (eg. Zp, Zs). Além disso, avaliou o desempenho de diferentes métodos de reamostragem de dados para balancear as classes da Formação Barra Velha. Com isso, foi possível aplicar a classificação às saídas da inversão sísmica, resultando em modelos 3D de fácies petroelásticas, unidades de fluxo e unidades geomecânicas. Os modelos apresentam acurácia média entre 70 e 95%. A distribuição das classes observadas nos modelos indicam maior proporção de fácies e unidades de rocha-reservatório nos intervalos de produção da Formação Barra Velha BVE100 e BVE300 e maior proporção de não-reservatório no BVE200. Essa distribuição está consistente com a interpretação geológica da área de estudo com base em estudos anteriores que identificaram, a partir de análise petrográfica, a maior presença de lama carbonática no BVE200, em contraste com os outros dois intervalos de produção. Além disso, os resultados de classificação indicam que os reservatórios ocorrem como estruturas alongadas na direção do alto estrutural (NE-SW), compartimentadas em corredores entre falhas que ocorrem na mesma direção. Os resultados demonstram o potencial do uso dessa ferramenta para classificação de dados sísmicos e caracterização de reservatórios do pré-sal.

Palavras-chave: Reservatórios de petróleo; Pré-sal; Inteligência artificial; Modelagem tridimensional.

ABSTRACT

Three-dimensional geological modeling of an oil field is fundamental for planning the stages of hydrocarbon exploration and production. Common outputs from reservoir modeling include 3D models of petroelastic facies, flow units, geomechanical units, and pore types. Typically, the understanding of the distribution of these subsurface classes is acquired through the integration of data obtained from seismic surveys and well logging. However, despite advances in determining these classes with some precision for each well, there is the problem of classification in the space between wells. Generally, this classification is done by arbitrarily marking homologous zones on scatter plots of elastic parameters, such as the P- and S-wave velocity ratio (Vp/Vs) and P-wave acoustic impedance (Zp). However, the interpreted region in the seismic volume critically depends on the size of the selected zone in the scatter plot, making the process very subjective. Therefore, the approach followed in this study involves the classification of units using a machine learning-based algorithm. This eliminates the need to arbitrarily define a threshold for separating homologous zones. The supervised classification algorithm was trained using part of the wells, with one well used for testing until all had been tested. The study evaluated the improvement in model performance by incorporating pseudo-curves extracted from seismic inversion products (e.g., Zp, Zs) into the training data. Additionally, it assessed the performance of different data resampling methods to balance the classes of the Barra Velha Formation. Consequently, the classification was applied to the outputs of the seismic inversion, resulting in 3D models of petroelastic facies, flow units, and geomechanical units. The models show an average accuracy between 70% and 95%. The distribution of classes observed in the models indicates a higher proportion of reservoir facies and units in the production intervals of the Barra Velha Formation BVE100 and BVE300, and a higher proportion of non-reservoir in BVE200. This distribution is consistent with the geological interpretation of the study area based on previous studies that identified, from petrographic analysis, a greater presence of carbonate mud in BVE200, in contrast to the other two production intervals. Furthermore, the classification results indicate that the reservoirs occur as elongated structures in the direction of the structural high (NE-SW), compartmentalized into corridors between faults that occur in the same direction. The results demonstrate the potential of using this tool for the classification of seismic data and the characterization of pre-salt reservoirs.

Keywords: Petroleum reservoirs; Pre-salt; Artificial intelligence; Three-dimensional modeling.

RESUMEN

La modelización geológica tridimensional de un campo petrolífero es fundamental para la planificación de las etapas de exploración y producción de hidrocarburos. Algunos productos comunes de la modelización de reservorios son los modelos 3D de facies petroelásticas, unidades de flujo, unidades geomecánicas y tipos de poros. Comúnmente, el entendimiento de la distribución de estas clases en el subsuelo se adquiere mediante la integración de datos obtenidos por levantamientos sísmicos y por la perfilación geofísica de pozos. Sin embargo, a pesar de los avances en la determinación de estas clases con cierta precisión para cada pozo, existe el problema de la clasificación en el espacio entre los pozos. En general, esta clasificación se realiza mediante la marcación arbitraria de zonas homologas en diagramas de dispersión de los parámetros elásticos, como la razón de velocidades de las ondas P y S (Vp/Vs) e impedancia acústica de la onda P (Zp). No obstante, la región interpretada en el volumen sísmico depende críticamente del tamaño de la zona seleccionada en el diagrama de dispersión, haciendo el proceso muy subjetivo. Dicho esto, el enfoque seguido en este estudio consiste en la clasificación de unidades a partir de un algoritmo basado en aprendizaje automático. De esta forma, se elimina la necesidad de definir arbitrariamente un umbral para la separación de zonas homologas. El algoritmo de clasificación supervisada fue entrenado a partir de parte de los pozos, utilizando 1 pozo para prueba hasta que todos hayan sido probados. El estudio evaluó la mejora del desempeño del modelo al incorporar a los datos de entrenamiento pseudo-curvas extraídas de los productos de inversión sísmica (e.g., Zp, Zs). Además, evaluó el desempeño de diferentes métodos de re-muestreo de datos para balancear las clases de la Formación Barra Velha. Con esto, fue posible aplicar la clasificación a las salidas de la inversión sísmica, resultando en modelos 3D de facies petroelásticas, unidades de flujo y unidades geomecánicas. Los modelos presentan una precisión media entre el 70% y el 95%. La distribución de las clases observadas en los modelos indica una mayor proporción de facies y unidades de roca-reservorio en los intervalos de producción de la Formación Barra Velha BVE100 y BVE300 y una mayor proporción de no-reservorio en el BVE200. Esta distribución es consistente con la interpretación geológica del área de estudio basada en estudios anteriores que identificaron, a partir de análisis petrográfico, una mayor presencia de lodo carbonático en el BVE200, en contraste con los otros dos intervalos de producción. Además, los resultados de la clasificación indican que los reservorios ocurren como estructuras alargadas en la dirección del alto estructural (NE-SW), compartimentadas en corredores entre fallas que ocurren en la misma dirección. Los resultados demuestran el potencial del uso de esta herramienta para la clasificación de datos sísmicos y la caracterización de reservorios del presal.

Palabras clave: Yacimientos de petróleo; Pre-sal; Inteligencia artificial; Modelado tridimensional.

SUMÁRIO

1. INTRODUÇÃO..... 11
2. OBJETIVOS..... 15
3. ÁREA DE ESTUDO..... 16
4. GEOLOGIA REGIONAL..... 17
4.1. Campo de Tupi..... 18
5. REVISÃO BIBLIOGRÁFICA..... 22
5.1. Fundamentos da Aprendizagem de Máquina..... 22
5.2. Modelos baseados em árvores de decisão..... 26
6. MATERIAIS..... 31
6.1. Perfis Geofísicos de Poços..... 31
6.2. Dados sísmicos (Ocean Bottom Nodes)..... 36
6.3. Softwares..... 37
7. METODOLOGIA..... 38
7.1. Avaliação dos resultados..... 41
8. RESULTADOS..... 43
9. DISCUSSÃO..... 57
10. CONCLUSÕES..... 65
REFERÊNCIAS..... 67

1. INTRODUÇÃO

Em qualquer desenvolvimento de um campo de petróleo ou gás natural, convencional ou não-convencional, o conhecimento da litologia é um alicerce fundamental para caracterização dos reservatórios. Esse conhecimento litológico não apenas ajuda a reduzir os riscos de perfuração de poços secos ou marginais, como também contribui para uma gestão sustentável dos recursos naturais.

Por sua vez, os diferentes tipos de rocha possuem propriedades petrofísicas e elásticas diferentes, tais como porosidade, permeabilidade e módulos elásticos, as quais podem ser usadas para definir fácies litológicas (litofácies) (Nieto *et al.*, 2013), fácies petroelásticas e unidades de fluxo. Enquanto as classes podem ser determinadas com certa precisão para cada poço, utilizando dados de perfis geofísicos integrados com amostras de testemunhos, há o problema da classificação dessas classes no espaço entre os poços (e.g. Bosch *et al.*, 2010; Zhao *et al.*, 2014).

Essa classificação constitui um desafio pois as variações espaciais das propriedades dos reservatórios e a correlação entre diferentes propriedades são complexas e, geralmente, não podem ser descritas por meio do uso de funções determinísticas simples (Doyen, 2007).

Em reservatórios carbonáticos, que representam a maior parte dos reservatórios de hidrocarbonetos no mundo, esse desafio é ainda maior. A caracterização sísmica desses reservatórios é dificultada devido às mudanças físicas, químicas e biológicas que envolvem a sedimentação e a diagênese pós-deposicional dos carbonatos, resultando em uma grande heterogeneidade de propriedades das rochas (Zhao *et al.*, 2014).

Neste contexto, esta pesquisa aplicou um fluxo de trabalho envolvendo a classificação de fácies petroelásticas, unidades de fluxo e unidades geomecânicas em dados de inversão sísmica de uma área do Campo de Tupi (antigo Campo de Lula). Este campo está localizado no polígono do Pré-Sal da Bacia de Santos, sendo sua rocha reservatório constituída de carbonatos da Formação Barra Velha (Wang *et al.*, 2013).

O Campo de Tupi conta com uma robusta base de dados de perfilagem geofísica de poços e de dados sísmicos de nós-de-fundo-oceânico (OBN, *Ocean Bottom Nodes*). Nesse

sentido, este trabalho consiste em uma integração desses conjuntos de dados. No geral, o problema básico abordado é: como relacionar os volumes de parâmetros elásticos, obtidos das inversões sísmicas, aos perfis geofísicos de poços?

Uma maneira rotineira de se fazer isso é utilizar diagramas de dispersão entre esses parâmetros para destacar certas zonas no volume sísmico, que correspondam às determinadas litofácies (Hampson, 2010). Entretanto, a área interpretada no volume sísmico depende criticamente do tamanho da zona selecionada no diagrama de dispersão (Hampson, 2010).

Nesse sentido, em vez de definir um limiar de corte arbitrário para delimitar a zona de interesse, a classificação bayesiana utiliza a teoria do campo aleatório para a caracterização de reservatórios (Doyen, 2007). Essa classificação é fundamentada no Teorema de Bayes, que descreve a probabilidade de um evento com base em um conhecimento a priori relacionado ao evento. Nesse contexto, os dados de poços são incorporados como conhecimento a priori, influenciando na probabilidade de ocorrência da fácie (ou unidade) no volume de dados sísmicos. Destacam-se os trabalhos de Penna & Lupinacci (2021) e Mattos et al. (2022) que aplicaram classificação bayesiana para obter modelos, respectivamente, de unidades de fluxo e fácies petroelásticas de carbonatos do pré-sal na Bacia de Santos.

Outra abordagem consiste na geoestatística, uma técnica estatística que permite a modelagem espacial de variáveis geológicas, enfatizando a importância das correlações espaciais na avaliação dos atributos (Matheron, 1963). Ao incorporar a estrutura espacial dos dados, a geoestatística fornece uma poderosa ferramenta para a caracterização de reservatórios, permitindo uma interpretação precisa das propriedades do reservatório (Pyrcz & Deutsch, 2014).

A aprendizagem de máquina, por sua vez, envolve o desenvolvimento de algoritmos capazes de aprender, a partir de dados, a fazer previsões ou classificações sem serem explicitamente programados para tarefas específicas (Bishop, 2006). No contexto da caracterização de reservatórios, a aprendizagem de máquina pode ser aplicada para identificar padrões complexos nos dados sísmicos e de poços, facilitando a classificação de fácies ou a predição de propriedades petrofísicas. Esta abordagem permite uma análise mais flexível e

profunda dos dados geológicos, contribuindo significativamente para a precisão na exploração de recursos naturais.

Nesse sentido, a abordagem seguida neste trabalho consiste na aplicação de um aprendizado supervisionado para que um algoritmo mapeie uma função que conecte os dados de entrada (perfis de poços) aos dados de saída (Fácies Petroelásticas, Unidades de Fluxo e Unidades Geomecânicas).

Os principais desafios da aplicação de aprendizagem de máquina em geociências incluem a escassez de dados rotulados, a complexidade e heterogeneidade dos dados geológicos, e a dificuldade em transferir modelos de aprendizado de máquina entre diferentes conjuntos de dados geocientíficos (Dramsch, 2019).

Dramsch et al. (2018c) demonstraram que a transferência de aprendizagem pode diminuir a necessidade de grandes quantidades de dados rotulados em rotinas de interpretação sísmica. Os autores aplicaram redes convolucionais (CNN, *Convolutional Neural Network*) com arquiteturas do estado da arte, pré-treinadas em imagens naturais, para interpretar seções sísmicas. O treinamento de redes neurais com imagens naturais envolve o reconhecimento de padrões visuais complexos. Além disso, este é um domínio rico em dados rotulados. Esse treinamento possibilita que a rede aprenda características genéricas de texturas, formas e contornos, as quais podem ser transferidas para tarefas em domínios específicos, como a interpretação sísmica. O estudo mostrou que as redes generalizáveis podem ser transferidas para dados sísmicos e superar redes menores treinadas do zero (Dramsch et al., 2018c).

Uma outra abordagem para lidar com os desafios da aprendizagem de máquina em geociências foi apresentada por Alvi (2023). O autor utilizou um conjunto de dados sintéticos para avaliar o desempenho do modelo ao incorporar vínculos (*constraints*) aos dados de treino, como informações sobre o tempo geológico (profundidade relativa). Além disso, o autor demonstrou que a modelagem de um reservatório complexo, com vários ambientes deposicionais, requer dados sísmicos de diferentes frequências para obter o melhor desempenho. Essa faixa de frequência depende da espessura e consistência das fácies (Alvi, 2023).

Alternativamente, este estudo aplica um fluxo de trabalho para geração de modelos geológicos 3D, que utiliza perfis geofísicos de poços para treinar um algoritmo para

classificar produtos de inversões sísmicas. Para lidar com a falta de balanceamento das classes, que se deve à heterogeneidade do reservatório, foram avaliadas as aplicações de algoritmos de reamostragem de dados (He & Garcia, 2009). Além disso, avaliou-se o desempenho do modelo ao ser treinado com pseudocurvas extraídas dos dados sísmicos visando vincular os dados de escala de poço aos dados de escala sísmica.

Dentre os principais algoritmos de aprendizagem de máquina aplicados em aprendizagem supervisionada estão os baseados em Árvores de Decisão, como o *Random Forest* e *Extreme Gradient Boosting* (XGBoost). O algoritmo XGBoost, proposto por Chen e Guestrin (2016), tem sido amplamente utilizado para resolver problemas de classificação e regressão, principalmente por causa de sua escalabilidade e desempenho (Ma *et al*., 2021). Portanto, neste trabalho o XGBoost foi aplicado para a classificação de fácies petroelásticas, unidades de fluxo e unidades geomecânicas visando a extrapolação desses dados da escala de poço para um volume sísmico.

2. OBJETIVOS

O objetivo principal do trabalho é desenvolver um fluxo de trabalho para obtenção de modelos tridimensionais de fácies petroelásticas, unidades de fluxo e unidades geomecânicas de uma área do Campo de Tupi, Bacia de Santos, através da aplicação de algoritmos baseados em aprendizagem de máquina. Para alcançar este objetivo principal, as seguintes metas principais deverão ser cumpridas:

- Desenvolver e treinar os modelos de classificação utilizando os dados de perfilagem geofísica de poço como dados de treino;
- Aplicar os modelos no volume inteiro onde há dados sísmicos disponíveis;
- Avaliar a qualidade dos modelos finais e interpretar geologicamente os resultados.

3. ÁREA DE ESTUDO

A área de estudo desta pesquisa consiste no Campo de Tupi, situado na porção nordeste da Bacia de Santos (Figura 1). Confirmado em 2006, o Campo de Tupi é o primeiro campo supergigante descoberto no território nacional (Petrobras, 2010). Essa denominação é dada a campos petrolíferos com volume de recuperação acima de 5 bilhões de barris de óleo equivalente (boe) (Ivanhoe & Leckie, 1993). O Campo de Tupi conta com um volume recuperável da ordem de 8.3 bilhões de boe e consiste, atualmente, no maior produtor de petróleo e gás natural do Brasil (Petrobras, 2010; ANP, 2024).

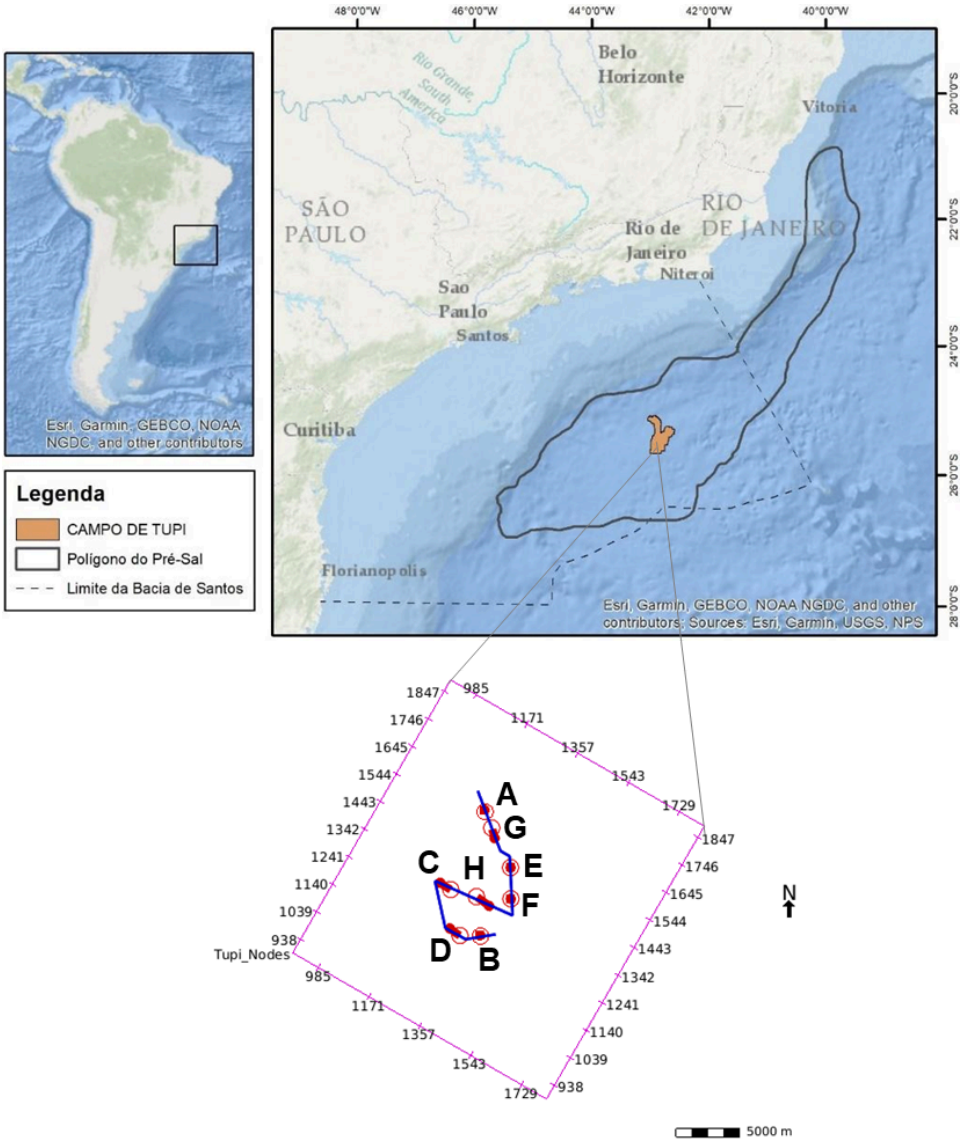

Figura 1: Localização da área de estudo, no polígono OBN (*Ocean Botton Nodes*) do Campo de Tupi, destacando os poços utilizados e a sessão sísmica apresentada nos resultados. Polígono do Pré-Sal e limite da Bacia de Santos: Riccomini *et al.* (2012). Localização do Campo de Tupi: ANP (2018). Basemap: ESRI.

4. GEOLOGIA REGIONAL

A Bacia de Santos está localizada na margem continental brasileira, na região sudeste, limitada pelos altos de Florianópolis e Cabo Frio (Moreira *et al.*, 2007). A bacia ocupa uma área de cerca de 352.000 km² e apresenta espessuras sedimentares superiores a 10 km nos principais depocentros (Chang *et al.*, 2008).

A sua formação é resultado da ruptura do supercontinente Gondwana, levando à separação das placas sul-americana e africana durante o Cretáceo Inferior (Chang *et al.*, 2008). Sua evolução tectono-sedimentar pode ser subdividida em três fases tectônicas: Rifte, Pós-Rifte e Drifte (Figura 2) (Moreira *et al.*, 2007).

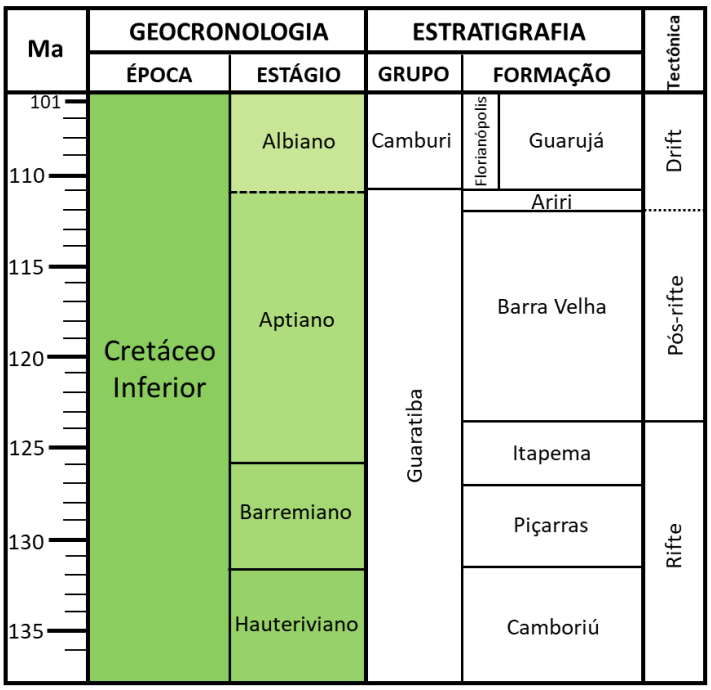

Figura 2: Carta estratigráfica do Cretáceo Inferior da Bacia de Santos. (Modificado de Moreira *et al.*, 2007).

A fase Rifte é caracterizada por esforços extensionais referente ao rifte Sul-Atlântico, que se propagou do Sul em direção ao Norte, resultando na separação das placas tectônicas Sul-americana e Africana (Chang *et al.*, 2008). O início desta fase é marcado por vulcanismo intenso registrado pelos basaltos da Formação Camboriú (Mohriak *et al.*, 2008). Com a evolução do sistema rifte, foi depositada a Formação Piçarras, caracterizada por arenitos e conglomerados de leques aluviais intercalados com folhelhos lacustres, siltitos e arenitos (Moreira *et al.*, 2007). Em seguida, ocorreu a deposição da Formação Itapema,

formada por sedimentos de leques aluviais e intercalação de calcirruditos e folhelhos escuros provenientes de incursões marinhas (Moreira *et al.*, 2007).

A fase Pós-rifte é registrada pelas Formações Barra Velha e Arri, que constituem o Grupo Guaratiba. (Moreira *et al.*, 2007). A Formação Barra Velha foi depositada em um ambiente de grandes lagos evaporíticos rasos e hiper-alcalinos (Wright & Barnett, 2015). Sobrepondo a Formação Barra Velha, está a Formação Ariri, composta por evaporitos em espessuras superiores a 2 km (Moreira *et al.*, 2007; Mohriak *et al.*, 2008).

A fase final, Drift, é marcada pela transição da crosta, de continental para oceânica (Mohriak, 2003). Moreira *et al.* (2007) associam a esta fase os grupos Camburi, Frade e Itamambuca.

4.1. Campo de Tupi

O Campo de Tupi, situado na porção nordeste da Bacia de Santos, se insere no sistema petrolífero Piçarras-Itapema/Barra Velha (Wang *et al.*, 2013). A rocha geradora do sistema consiste na Formação Piçarras constituída de folhelhos de origem lacustre, depositados no contexto da metade superior da sequência Rifte. A rocha reservatório é constituída por carbonatos das Formações Itapema e Barra Velha. A armadilha do sistema é do tipo litológico-estrutural. Nesse sentido, ocorrem evaporitos da Formação Ariri como componente litológica da trapa, além da componente estrutural controlada por horsts desenvolvidos durante a fase rifte (Wang *et al.*, 2013).

A principal rocha reservatório do Campo de Tupi consiste nos carbonatos da Formação Barra Velha (Wang *et al.*, 2013). Gomes *et al.* (2020) propõem que as fácies da Formação Barra Velha são formadas por três componentes básicos: lama carbonática, esferulitos e *shrubs*.

A lama carbonática é caracterizada como um material de granulação fina que inclui argilo-minerais, calcita, dolomita e sílica (Gomes *et al.*, 2020). Os esferulitos se caracterizam como agregados de calcita esféricos a subesféricos com extinção radial, comumente dolomitizados ou recristalizados (Gomes *et al.*, 2020). Os schrubs consistem em

cristais de calcita fibrosos e de granulação muito grossa com formato de arbustos (Gomes *et al.*, 2020).

Com base na proporção desses elementos básicos, Gomes et al. (2020) propuseram um esquema de classificação de 17 fácies para as rochas da Formação Barra Velha: *spherulitestones*, *shrubby spherulitestones*, *spherulitic shrubstones*, *shrubstones*, *spherulitic shrubstones with mud*, *shrubby spherulitestones with mud*, *muddy spherulitestones*, *spherulitic mudstones*, *mudstone*, *wackestones*, *packstones*, *grainstones*, *Mg-clay mudstone*, *calcimudstone*, *dolomudstone*, *mixed mudstone* e *siliceous mudstone* (Figura 3). O esquema proposto leva em consideração a contribuição relativa de componentes deposicionais, diagenéticos e mineralógicos.

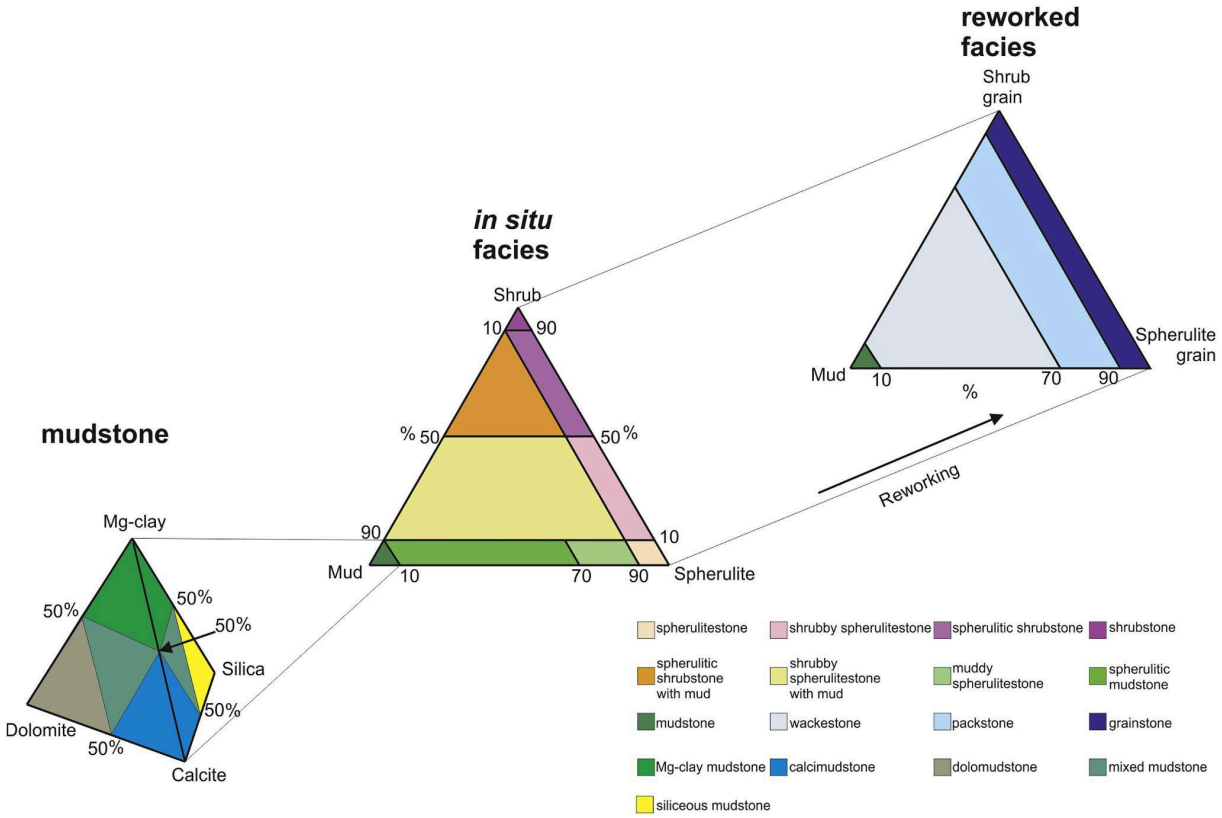

Figura 3: Esquema de classificação de fácies da Formação Barra Velha, baseado em 3 diagramas triangulares. Fonte: Gomes et al. (2020).

Rebelo et al. (2022) agruparam as fácies da Formação Barra Velha definidas por Gomes et al. (2020) em 6 associações de fácies de acordo com suas características

deposicionais: *Deep Lake*, *Lower Shoreface*, *Upper Shoreface*, *Lower Foreshore*, *Upper Foreshore* e *Backshore* (Figura 4).

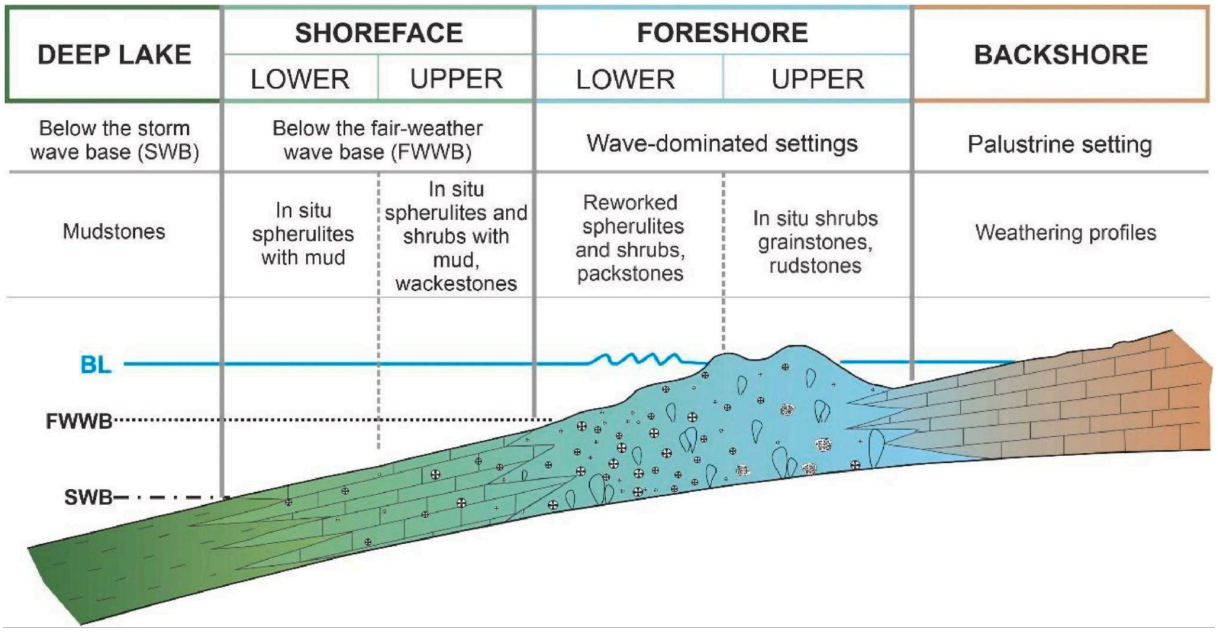

Figura 4: Esquema de distribuição de associações de fácies da Formação Barra Velha. BL: *Base level* (Nível de Base); FWWB: *Fair-weather wave base* (Nível de ondas de tempo bom); e SWB: *Storm wave base* (Nível de ondas de tempestade). Fonte: Rebelo et al. (2022).

A *Deep Lake* está associada à fácies *mudstone*. Indica deposição em ambientes profundos abaixo do nível da onda de tempestade (Rebelo et al. 2022).

As associações *Shoreface* foram depositadas em ambientes localizados abaixo da linha de ondas de tempo bom (*fair-weather wave base*, FWWB). A *Lower Shoreface* inclui as fácies *muddy spherulitestones* e *spherulitic mudstones* (Rebelo et al., 2022). Os principais elementos desta associação são a lama carbonática, argilas ricas em Al e Mg e esferulitos. A *Upper Shoreface* inclui *shrubby spherulitestones with mud*, *spherulitic shrubstones with mud* e *wackestones*, sendo caracterizada pela presença de fácies in-situ e retrabalhadas, com quantidade significativa de argila (Rebelo et al., 2022).

As associações *Foreshore* representam um ambiente mais raso, com alguma influência de ondas nas margens do lago, o que levaram à formação de fácies *in situ* e retrabalhadas (Rebelo et al., 2022). A *Lower Foreshore* inclui *shrubby spherulitestones*,

spherulitic shrubstones, *spherulitestones*, e *packstones* e pode conter uma quantidade limitada de argila. A *Upper Foreshore* inclui *grainstones* e *shrubstones*, depositadas em águas bem rasas (Rebelo et al., 2022).

A *Backshore* compreende os perfis de alteração. Rebelo et al. (2022) interpretam essa associação como um ambiente palustre, com modificações dos carbonatos por processos biogênicos e exposição subaérea, levando à formação de carbonatos pedogenéticos e calcretes.

5. REVISÃO BIBLIOGRÁFICA

Este capítulo visa apresentar fundamentos teóricos acerca da metodologia aplicada nesta pesquisa. Nesse sentido, serão apresentados conceitos sobre aprendizagem de máquina e algoritmos baseados em árvores de decisões.

5.1. Fundamentos da Aprendizagem de Máquina

Witten *et al*. (2005) definem a aprendizagem de máquina como o processo em que uma máquina muda o seu comportamento a fim de melhorar sua performance no futuro. Os algoritmos baseados em aprendizagem de máquina são amplamente empregados para a resolução de problemas de regressão (predição de um dado numérico) e de classificação (inferência de um dado categórico). A aprendizagem pode ser supervisionada ou não-supervisionada.

A aprendizagem não supervisionada consiste no processo de aprendizagem em que não há resposta correta e não há professores (Mahesh, 2020). O algoritmo não supervisionado aprende alguns atributos a partir do conjunto de dados de entrada e reconhece a classe dos novos dados inseridos no modelo a partir dos atributos aprendidos anteriormente (Figura 5). Os algoritmos de aprendizagem não supervisionada são frequentemente aplicados para solucionar problemas de agrupamento de dados e de redução de atributos.

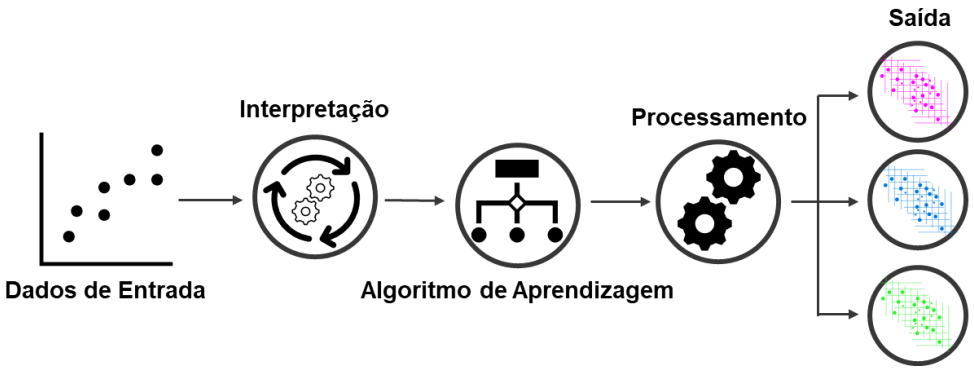

Figura 5: Fluxograma típico de uma aprendizagem não supervisionada. Adaptado de Mahesh (2020).

Na aprendizagem supervisionada, por outro lado, os dados de entrada são divididos em dados de treino e dados de teste, conforme a Figura 6. O algoritmo tem a tarefa de aprender uma função que mapeie a relação entre uma entrada (parâmetros, ex.: Zp, Zs,

Vp/Vs, RHOB) com uma saída (ex: Fácies Petroelásticas), aprendendo padrões com base no conjunto de dados de treino e aplicando esses padrões para predição ou classificação (Mahesh, 2020).

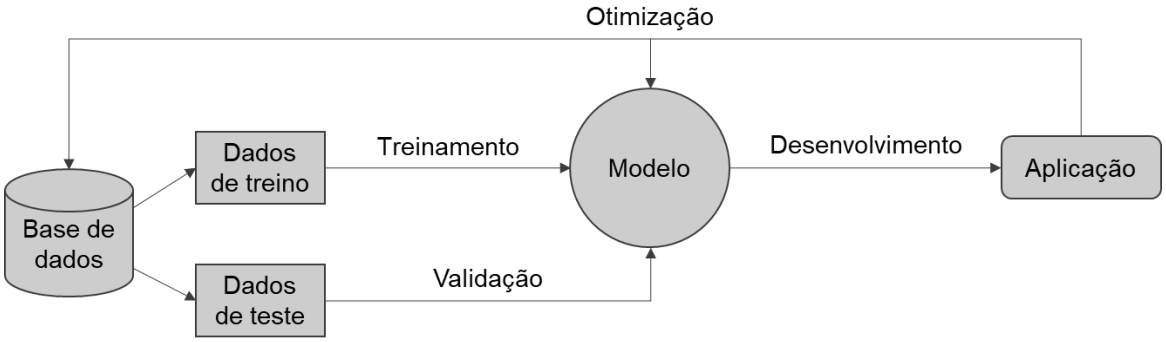

Figura 6: Fluxograma típico de uma aprendizagem supervisionada. Adaptado de Mahesh (2020).

Duas grandezas estatísticas normalmente utilizadas para avaliar os resultados de métodos de aprendizagem de máquina são o viés (*bias*) (Equação 5.1) e a variância (Eq. 5.2). O viés consiste em um erro gerado a partir de premissas incorretas assumidas pelo modelo, tal como assumir a linearidade de uma função interpoladora em vez de uma equação de maior grau (Skiena, 2017). É calculado pela diferença entre o valor esperado ($E_{\theta}(\hat{\theta})$) e o calculado (θ) (Eq. 5.1).

$$\text{Viés} = E_{\theta}(\hat{\theta}) - \theta \qquad (5.1)$$

$$\text{Variância} = \frac{1}{m-1}\sum_{i=1}^{m}(x_i - \bar{x})^2 \qquad (5.2)$$

O viés pode ser interpretado como a incapacidade do método em capturar a relação verdadeira entre as variáveis. Isso ocorre na Figura 7.A, em que um modelo representado pela reta vermelha não captura corretamente a relação entre as variáveis A e B, indicada pela curva azul. Erros de viés produzem modelos com *underfitting*, que ocorrem quando os modelos não se ajustam adequadamente aos dados de treino (Skiena, 2017).

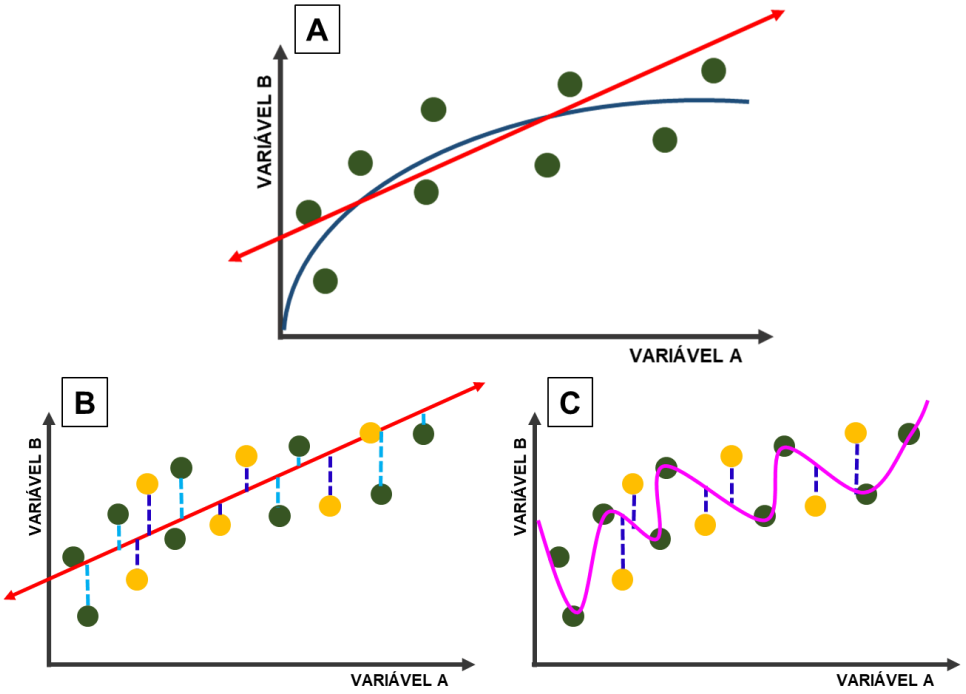

Figura 7: Gráficos de dispersão indicando a relação entre as variáveis A e B. Os pontos verdes representam dados de treino e os pontos laranjas representam dados de teste. **7.A**) A curva azul indica a relação real entre as variáveis e a reta vermelha indica um modelo com viés. **7.B**) A linha tracejada indica a distância entre o valor real e o valor previsto pelo modelo representado pela reta vermelha. A linha azul claro representa a distância para os dados de treino e a linha azul escuro representa a distância para os dados de teste. Como as adições das linhas azul claro e das linhas azul escuro resultam em valores próximos, a reta vermelha representa um modelo com baixa variância. **7.C**) Exemplo de um modelo com alta variância (*overfitting*).

A variância é um erro associado à sensibilidade do modelo em relação às flutuações no conjunto de dados de treino (linha tracejada azul claro da Fig. 7.B) e de teste (linha tracejada azul escuro da Fig. 7.B), correspondendo à diferença entre o valor calculado para cada conjunto de dados (Skiena, 2017).

A variância pode ser calculada pela Equação 5.2, na qual *m* corresponde ao total de amostras, x_i ao valor da amostra *i* e $\bar{x}$ ao valor médio das amostras, sendo que no caso da quantidade de amostras ser muito grande, a diferença entre o valor de *m* e *m-1* torna-se irrelevante e utiliza-se *m* no lugar de *m-1* (Igual & Seguí, 2017).

Erros de variância indicam modelos com *overfitting*, que ocorrem quando o modelo se ajusta excessivamente aos dados de treino, que contém ruídos. Devido ao ajuste do modelo aos ruídos dos dados de treino, o mesmo apresenta um ótimo desempenho quando

aplicado aos dados de treino, porém a acurácia é reduzida radicalmente quando aplicado aos dados de teste (Skiena, 2017).

Modelos baseados em premissas ou princípios tendem a apresentar maior viés enquanto que modelos dirigidos por dados são mais suscetíveis à overfitting (Skiena, 2017). Os algoritmos ideais possuem baixo viés, modelando a relação entre as variáveis com acurácia, e baixa variabilidade, apresentando resultados consistentes entre diferentes conjuntos de dados.

O modelo, em aprendizagem supervisionada, se refere à estrutura matemática por meio da qual uma predição y_i é feita a partir de um conjunto de dados de entrada x_i. Como exemplo, um modelo linear $\hat{y}_i = \Sigma_j \theta_j x_{ij}$ que consiste em uma combinação linear dos pesos dos atributos de entrada (XGBoost Documentation, 2022). Os parâmetros consistem na porção não determinada do modelo, que deve ser aprendida a partir dos dados de treino. Em modelos de regressão linear, os parâmetros consistem nos coeficientes θ.

A tarefa de treinamento do modelo envolve encontrar os parâmetros que ajustam melhor os dados de treino (x_i) aos rótulos (y_i) (XGBoost Doc., 2022). Nesse sentido, a função objetivo consiste em uma função que indica o quanto o modelo se ajusta aos dados de treino. As funções objetivo (Equação 5.3) são constituídas de duas partes: perda de treinamento (*training loss*, $L(\theta)$) e termo de regularização (*regularization term*, $\Omega(\theta)$).

$$obj(\theta) = L(\theta) + \Omega(\theta) \qquad (5.3)$$

A perda de treinamento mede o quanto o modelo é preditivo com relação ao conjunto de dados de treino, sendo que uma escolha comum para *L* é o erro quadrático médio (*Mean Squared Error, MSE*, Equação 5.4). O termo de regularização controla a complexidade do modelo, ajudando a evitar o overfitting.

$$L(\theta) = \sum_i (y_i - \hat{y}_i)^2 \qquad (5.4)$$

A relação entre as funções de perda de treinamento e termo de regularização são apresentadas na Figura 8. A Fig. 8.A exibe os dados de interesse do usuário em função do

tempo (t). A Figura 8.B apresenta um modelo com número de divisões alto, indicando um termo de regularização ($\Omega(f)$) muito alto, assim como a complexidade do modelo. A Figura 8.C indica um modelo cujo ponto de divisão foi feito em um valor não apropriado, com isso, apesar de manter a baixa complexidade, o modelo não se ajusta adequadamente aos dados de treino e, portanto, apresenta um valor de perda de treinamento ($L(f)$) alto. Finalmente, a Figura 8.D indica um modelo com um bom balanço entre complexidade $\Omega(f)$ e perda de treinamento $L(f)$. O princípio geral consiste na busca por um modelo simples e preditivo, com um balanço controlado entre viés e variância (XGBoost Doc., 2022).

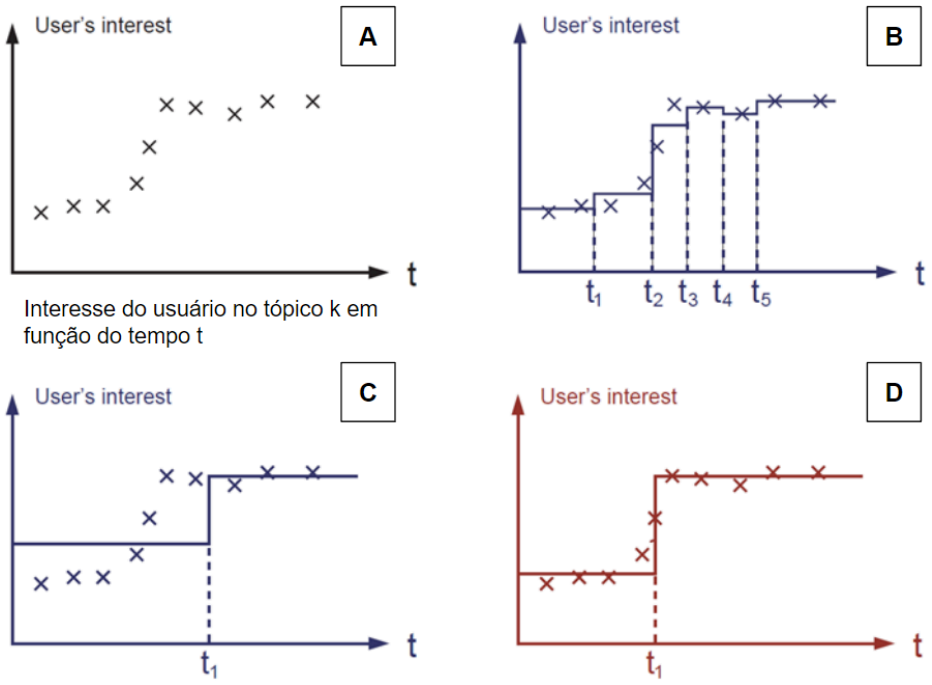

Figura 8: Relação entre perda de treinamento ($L(f)$) e termo de regularização ($\Omega(f)$). **A**) Conjunto de dados indicando interesse em função do tempo. **B**) Modelo com valor alto de $\Omega(f)$. **C**) Modelo com valor alto de $L(f)$. **D**) Modelo com um balanço bom entre $\Omega(f)$ e $L(f)$. Adaptado de XGBoost Doc. (2022).

5.2. Modelos baseados em árvores de decisão

Dentre os principais algoritmos baseados em aprendizagem supervisionada estão as árvores de decisão. Árvores de decisão são gráficos que representam escolhas e seus resultados na forma de uma árvore (Figura 9).

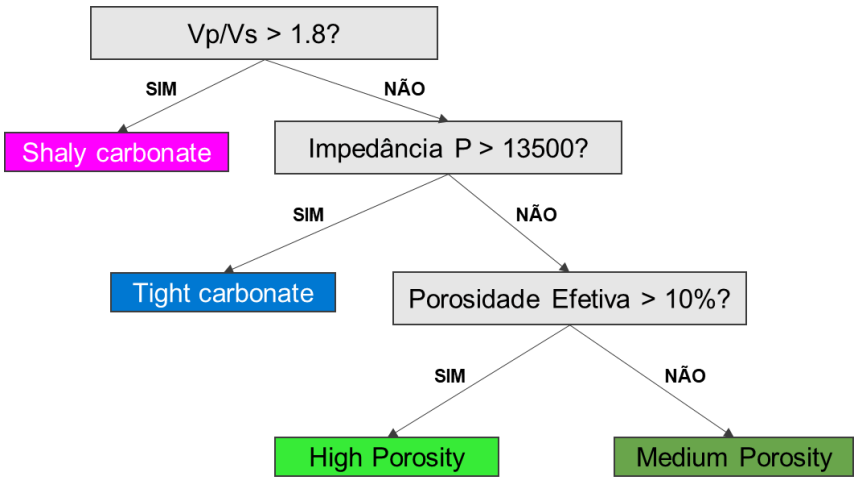

Figura 9: Árvore de decisão para classificação de fácies petroelásticas com base nos valores de corte estabelecidos por Mattos *et al*. (2022) para dados geofísicos de poços.

Diversos algoritmos (e.g. *Random Forest, AdaBoost, Gradient Boosting, XGBoost*) trabalham com conjuntos de árvores de decisão (*decision tree ensembles*). Os modelos de conjuntos de árvores são constituídos de uma série de árvores de classificação e regressão (*Classification and regression trees, CART*) (XGBoost Doc., 2022). Cada folha da CART conta com uma pontuação. Nesse sentido, a predição se dá pela soma da pontuação da folha de cada árvore (Equação 5.5, onde onde K é o número de árvores, f_k é uma função no espaço funcional, e Γ é o conjunto de todos os CARTs possíveis). A Equação 5.6 apresenta a função objetivo a ser otimizada.

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i),\ f_k \in \Gamma \qquad (5.5)$$

$$obj = \sum_{i}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} w(f_k) \qquad (5.6)$$

Árvores de decisão construídas a partir de bases de dados muito extensas tendem a formar estruturas complexas, com grandes profundidades de árvores e número de folhas. Conforme o número de árvores que compõem o modelo aumenta, também cresce a taxa de erro de generalização que converge até um dado limite (Breiman, 2001). Nesse sentido, alguns métodos são empregados visando melhorar a generalização dos algoritmos baseados em árvores de decisão. Dentre esses métodos destacam-se o *Bootstrap*, *Aggregate, Bagging e Boosting*.

O *bootstrap* consiste na técnica de utilizar somente uma parte das amostras e das variáveis disponíveis para a construção de um conjunto de árvores de decisão. Dessa forma, o algoritmo constrói árvores menos complexas que são capazes de fazer melhores generalizações. Adicionalmente, alguns modelos utilizam o *aggregate* para tomar decisões em problemas de regressão e classificação. Esta técnica consiste na classificação de cada amostra do conjunto de dados para cada árvore de decisão criada, de modo que a classificação final da amostra é a opção mais votada entre as árvores da floresta. A utilização das técnicas *bootstrap* e *aggregate* em conjunto é chamada de *bagging* (Breiman, 1996).

O *Random Forest* (Breiman, 2001) é um algoritmo que utiliza o *bagging* para construir uma floresta com um número pré-definido de árvores de decisão de profundidade máxima não fixada que são utilizadas para classificar ou prever dados. No *Random Forest* as árvores são criadas de forma independente uma das outras e possuem o mesmo peso na votação para a classificação final.

Em contraste, o *boosting* cria árvores de forma sequencial. O *boosting* se baseia na ideia de que a criação de várias regras para classificação, ou predição, que geram resultados de baixa acurácia é mais rápida e gera melhores resultados que a criação de uma única regra muito complexa (Schapire, 2003). O treinamento do modelo, como em todo aprendizado supervisionado, se dá a partir da otimização da função objetivo do modelo (XGBoost Doc., 2022).

A Equação 5.7 apresenta a função objetivo a ser otimizada do modelo representado na Equação 5.6. A variável $\hat{y}_i^{(t)}$ representa o valor da predição no passo *t* (Equação 5.8). Com isso, a cada etapa o modelo corrige o que aprendeu e adiciona uma nova árvore, de forma sequencial, que otimiza a função objetivo (Equação 5.9) (XGBoost Doc., 2022).

$$obj^{(t)} = \sum_{i}^{n} l(y_i, \hat{y}_i^{(t)}) + \sum_{k=1}^{K} w(f_k) \tag{5.7}$$

$$\hat{y}_i^{(t)} = \sum_{k=1}^{t} f_k(x_i) = \hat{y}_i^{(t-1)} + f_t(x_i) \tag{5.8}$$

$$obj^{(t)} = \sum_{i}^{n} l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + w(f_t) + Const. \quad (5.9)$$

O *Adaptative Boosting (AdaBoost)* (Freund & Schapire, 1997) cria uma floresta de "*stumps*", que são árvores de decisão de profundidade igual a 1, ou seja, formada apenas por um nó e duas folhas. Em geral, *stumps* são classificadores limitados, gerando resultados de baixa acurácia. Para contornar isso, no *AdaBoost,* os *stumps* com menores taxas de erro possuem maior peso na votação. Além disso, cada *stump* é criado tendo como base os erros de classificação do *stump* criado anteriormente.

O *Gradient Boost* (Friedman, 2001) consiste em um método baseado em árvores de decisão em que as árvores são elaboradas a partir das folhas. O algoritmo atribui um valor (no caso de problemas de regressão) ou uma classe (em caso de problemas de classificação) para a folha e a completa com uma árvore de decisão com valores de profundidade e de número de folhas definidos previamente. Em seguida são geradas novas árvores, levando-se em consideração os erros da árvore anterior, assim como no método *AdaBoost*, com a diferença de que as árvores podem ser maiores que os *stumps* utilizados no *AdaBoost*.

Nesse sentido, o desempenho de cada árvore é mensurado através do resíduo em relação à diferença entre o previsto e o real para cada amostra (Friedman, 2001). A cada passo do *boosting*, é calculado o índice de saída de cada folha que é aplicado para corrigir os erros da classificação anterior e é utilizado para calcular o resíduo e a probabilidade de acerto da classificação. Um importante hiperparâmetro desse modelo é a taxa de aprendizagem (*learning rate*) que é utilizada para controlar a variância do modelo. A taxa de aprendizagem consiste em um peso (de 0 a 1) que indica a porcentagem em que a correção do resíduo será aplicada a cada passo do *boosting*. Em geral, esse parâmetro é mantido em um valor baixo (e.g. 0,3) visando a redução da variância.

O *XGBoost* (Chen e Guestrin, 2016) é um algoritmo baseado em árvores de decisão que utiliza a técnica *Gradient Boosting* para a solução de problemas de classificação e regressão. Nesse sentido, o algoritmo cria árvores de forma sequencial e considera os erros da árvore criada anteriormente, ajustando o modelo a cada passo do *boosting*. Para controlar a variância do modelo, o XGBoost conta com hiperparâmetros que controlam a construção das árvores (e.g. *learning rate*, *subsample*). A função objetivo do XGBoost pode ser representada

pela Equação 5.10, sendo formada por um termo de perda de treinamento (primeiro termo entre colchetes) e um termo de regularização (segundo termo entre colchetes). O XGBoost utiliza o erro médio quadrático (*Mean Squared Error, MSE*) como função de perda de treinamento.

$$obj^{(t)} = \sum_{i}^{n}[g_i f_t(x_i) + \frac{1}{2}h_i f_t^2(x_i)] + [\gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2] \quad (5.10)$$

6. MATERIAIS

A base de dados utilizada neste trabalho consiste em perfis geofísicos de poços e dados sísmicos 3D de nós de fundo do oceano (OBN, sigla do termo *Ocean Bottom Nodes*) (Cruz *et al*., 2021). Os dados foram fornecidos pela Agência Nacional de Petróleo, Gás Natural e Biocombustíveis (ANP).

6.1. Perfis Geofísicos de Poços

Os perfis geofísicos convencionais utilizados consistem em curvas de densidade volumétrica da rocha (RHOB), velocidades de onda compressional (Vp) e de cisalhamento (Vs). Esses dados foram utilizados para o cálculo da impedância da onda P (Zp), impedância da onda S (Zs), e razão Vp/Vs. Também foi utilizado neste trabalho o atributo CSI, que corresponde ao indicador Carbonato-Folhelho calculado a partir dos dados Zp e Zs (Mattos et al., 2022). A Figura 10 apresenta um exemplo do conjunto de dados de poços.

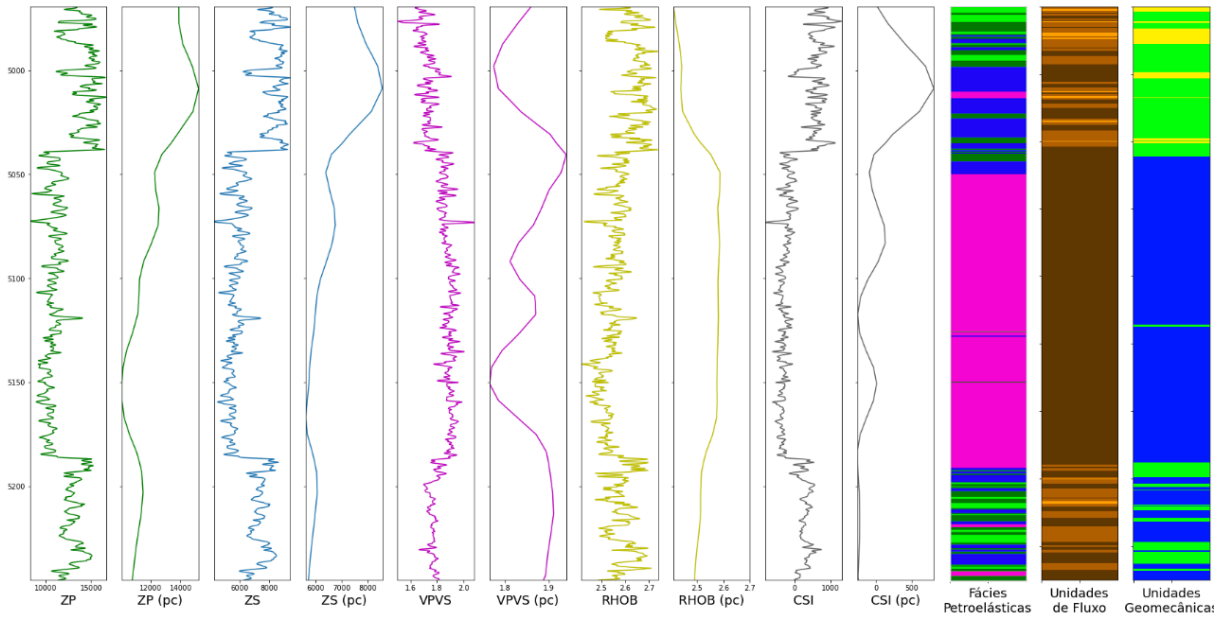

Figura 10: Exemplo de conjunto de dados de poço utilizados para o treinamento do modelo (Poço E).

As velocidades sísmicas (e.g. Vp, Vs) das formações geológicas investigadas, em escala de poço, são obtidas por perfis sônicos, também chamados de perfis acústicos (Kearey

et al., 2002). A aquisição deste conjunto de dados é feita por meio de uma sonda com dois receptores e uma fonte acústica. Essa sonda gera pulsos ultrassônicos a uma frequência entre 20 e 40 kHz. Dentro do poço, parte do pulso é refratado da parede da rocha para os receptores. Isso acontece porque a velocidade da rocha é maior que a do fluido de perfuração que preenche o poço. Com isso, a velocidade é determinada a partir da medida do tempo diferencial de trânsito entre os receptores (Kearey et al., 2002).

Os perfis de Densidade de Raios Gama (RHOB) são obtidos através da aplicação de uma fonte radioativa (e.g. Césio 137) na parede do poço. A fonte emite raios gama que interagem com os elétrons dos átomos da formação e perdem energia ao colidirem, fenômeno conhecido como espalhamento de Compton (Kearey et al., 2002). O número de colisões é dependente da abundância de elétrons presentes, que aumenta conforme a densidade da formação. Nesse sentido, a densidade é estimada pela proporção de radiação gama que retorna para o detector (Kearey et al., 2002).

A partir dos dados de velocidade das ondas P (Vp) e S (Vs), e da densidade (RHOB), foi possível o cálculo da impedância das ondas P (Zp) e S (Zs). Mattos et al (2022) utilizaram os valores de Zp e Zs para o cálculo do atributo CSI (*Carbonate-Shale Indicator*).

O CSI é um método semi-quantitativo aplicado a dados de poços e produtos de inversão sísmica para identificar a distribuição de litologias com maior presença de carbonatos e litologias mais argilosas (Mattos et al., 2022). O método se baseia na Impedância de Poisson, e incorpora o coeficiente de Poisson e a densidade das rochas em sua determinação gráfica. Por meio de diagramas de dispersão, é possível identificar os agrupamentos de dados que separam os carbonatos e as litologias argilosas. Essa divisão considera uma rotação ótima do eixo, expressa pelo termo c na equação 6.1.

Equação 6.1: $CSI = Zp - cZs$

Os perfis utilizados nesta pesquisa apresentam diferentes taxas de amostragem. Nesse sentido, foi aplicado um pré-processamento utilizando-se a técnica linear de reamostragem com taxa fixa de 0.1524 m (1 ft). O pré-processamento foi realizado pela equipe do Laboratório de Geofísica do IG-Unicamp por meio do software Techlog (Schlumberger®).

A base de dados também inclui as classificações, em escala de poço, de fácies petroelásticas (Mattos et al., 2022), unidades de fluxo (Rebelo et al., 2021), e unidades geomecânicas (Rojas, 2023). A Tabela 1 indica as classificações disponíveis para cada poço.

Tabela 1: Classificações disponíveis para cada poço utilizado neste trabalho.

Poços	Facies Petroelásticas	Unidades de Fluxo	Unidades Geomecânicas
A	x	x	x
B	x		x
C	x		x
D	x		x
E	x	x	x
F	x	x	x
G	x	x	x
H	x	x	x

Mattos *et al*. (2022) classificaram os carbonatos da Formação Barra Velha em 4 fácies petroelásticas: Carbonatos Fechados (Tight), Carbonatos Argilosos (Shaly), Carbonatos de Porosidade Média (Medium Porosity) e Carbonatos de Alta Porosidade (High Porosity) (Figura 11). A classificação é baseada em 4 parâmetros: Porosidade Efetiva, Razão Vp/Vs e Impedâncias das ondas P e S (Tabela 2). Os reservatórios abrangem as fácies de carbonatos com porosidade média e alta, enquanto que os carbonatos fechados e argilosos são considerados não reservatórios (Mattos et al., 2022).

Tabela 2: Valores de corte utilizados para determinar as fácies petroelásticas (Mattos *et al*., 2022).

Fácie	Porosidade Efetiva	Razão Vp/Vs	Impedância-P ($g/cm^3*m/s$)	Impedância-S ($g/cm^3*m/s$)
Fechado	< 7%	-	> 13500	> 7000
Argiloso	< 6%	>1.8	< 13500	< 7500
Porosidade Média	> 6% e <10%	< 1.8	12000 - 13500	-
Porosidade Alta	> 10%	< 1.8	10000 - 13000	-

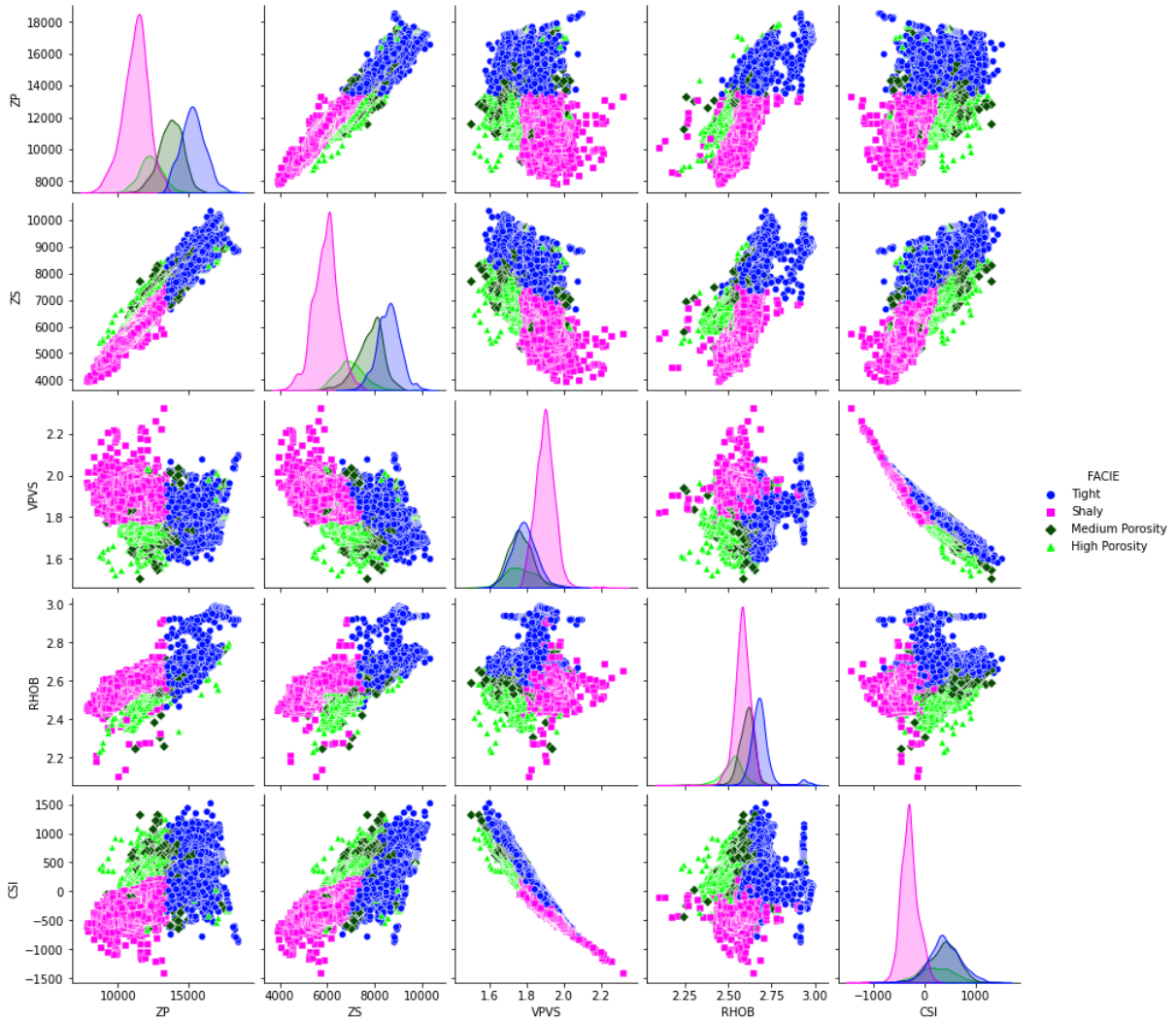

Figura 11: Diagramas de dispersão das variáveis densidade (RHOB), razão Vp/VS (VPVS), Impedância-P (ZP), Impedância-S (ZS) e CSI em função da classificação das fácies petroelásticas.

Rebello *et al*. (2022) classificaram as rochas da Formação Barra Velha em unidades de fluxo com base nos dados de *routine core analysis* (RCA). Os dados compreendem medidas de porosidade e permeabilidade de plugs e testemunhos de sondagem e dados de porosidade obtidos por ressonância magnética nuclear (NMR).

Nesse sentido, Rebello *et al*. (2022) utilizaram a classificação de Aguilera R35 (Equação 6.1), em que R35 é o raio da garganta de poro a uma saturação de 35% de mercúrio no teste de pressão de capilaridade, k é a permeabilidade (mD) e φ é a porosidade (%). Dessa forma, Rebello *et al*. (2022) definiram 4 unidades de fluxo, na seguinte ordem crescente de qualidade de reservatório: UF 1, UF 2, UF 3 e UF 4 (Figura 12).

$$R_{35} = 2,665\left(\frac{k}{\varphi}\right)^{0.45} \qquad (6.1)$$

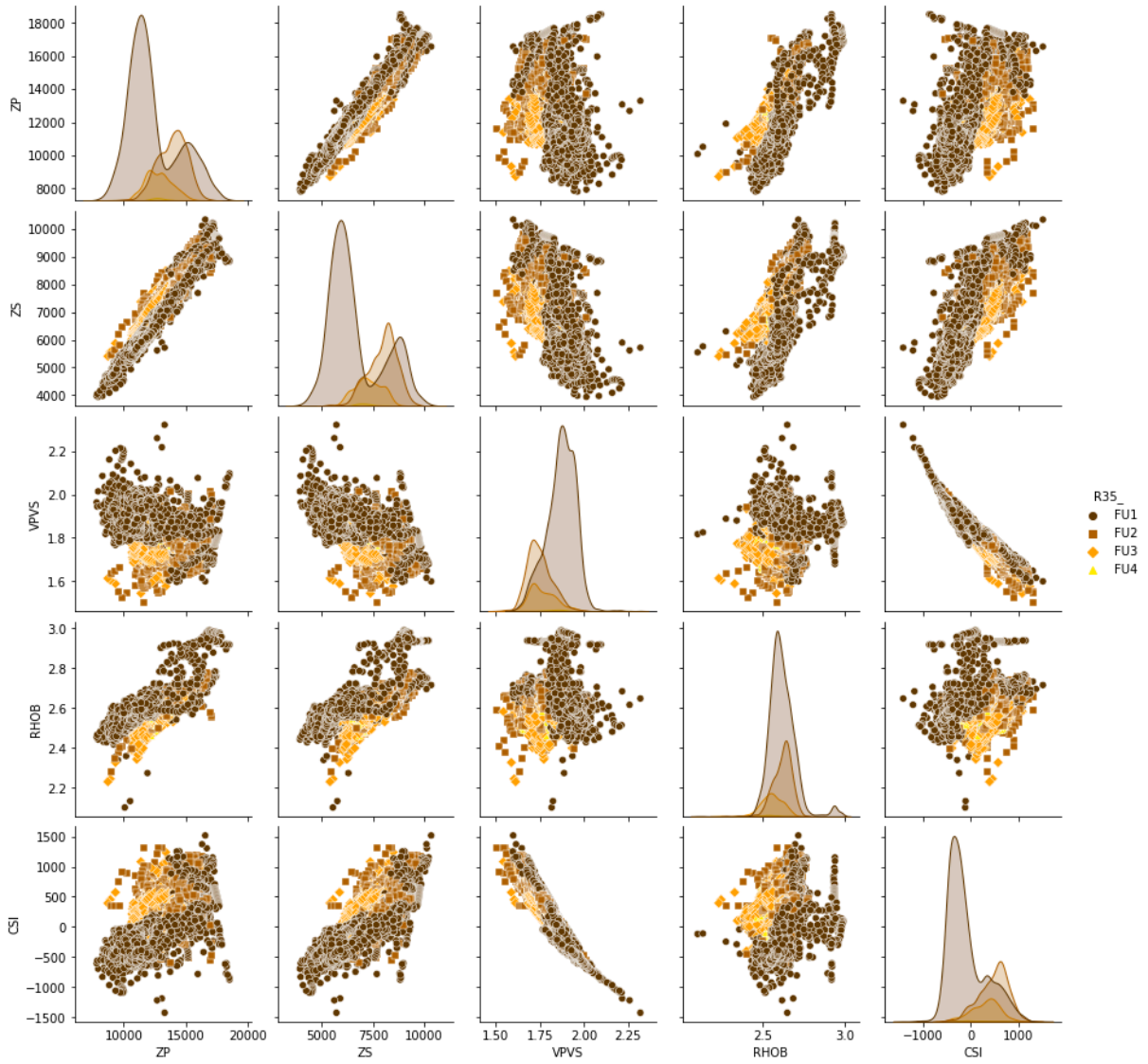

Figura 12: Diagramas de dispersão das variáveis densidade (RHOB), razão Vp/VS (VPVS), Impedância-P (ZP), Impedância-S (ZS) e CSI em função da classificação das unidades de fluxo.

Rojas (2023) definiu as unidades geomecânicas por meio de uma classificação não supervisionada utilizando os seguintes dados de entradas: Propriedades elásticas (E_s e σ_s), Resistência mecânica da rocha (UCS), critérios de falha (Φ_{Atrito}, C e CP), pressão do poro (P_{pore}), Tensão horizontal (SH_{Max} e SH_{Min}) e porosidade total (Φ_{Total}).

Com isso, Rojas (2023) definiu 3 Unidades Geomecânicas, GMU 1, com baixa UCS e alta Φ_{Total}, GMU 2, com valores médios de UCS e Φ_{Total} e GMU 3, com alta UCS e baixa Φ_{Total} (Figura 13).

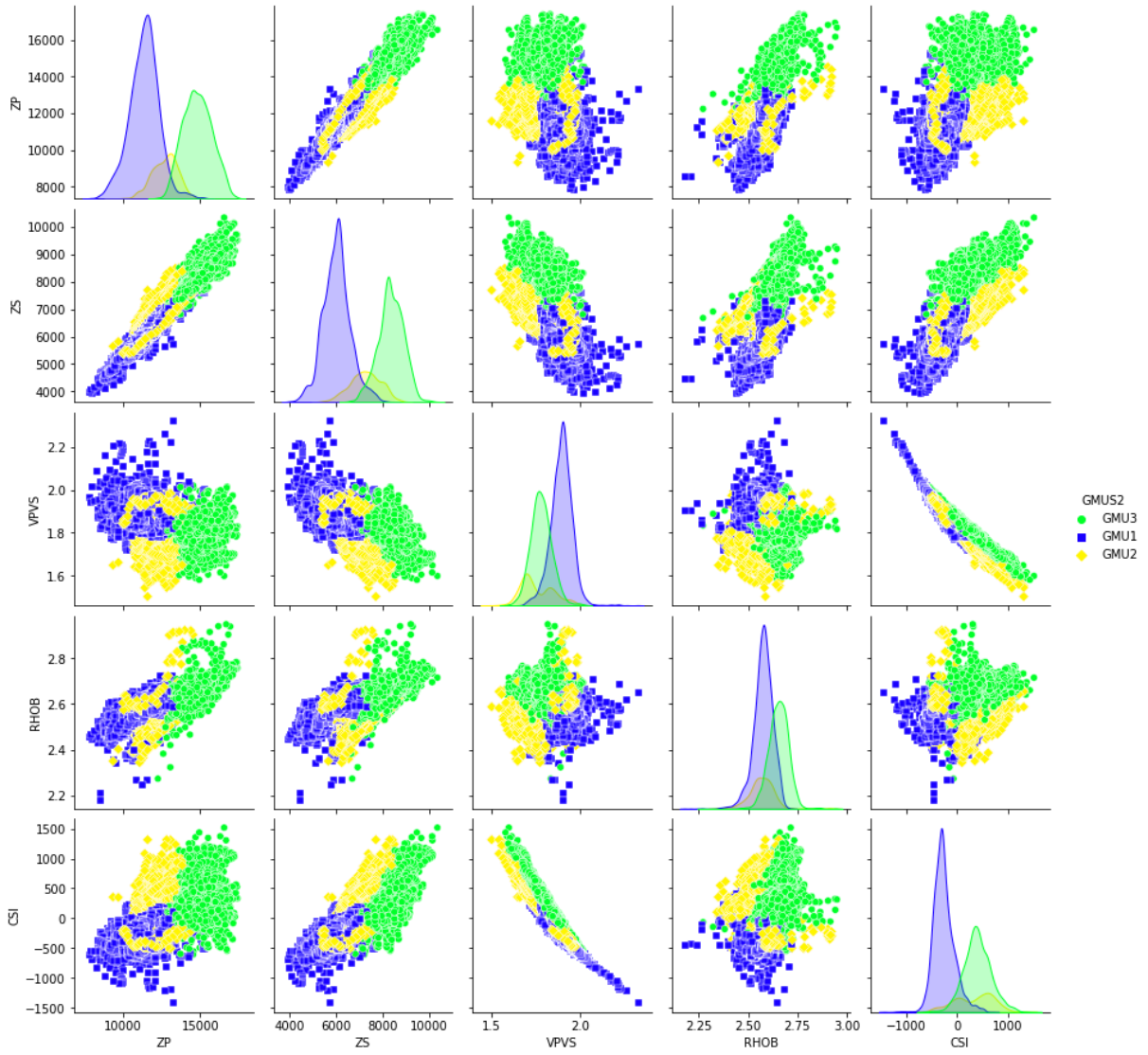

Figura 13: Diagramas de dispersão das variáveis densidade (RHOB), razão Vp/VS (VPVS), Impedância-P (ZP), Impedância-S (ZS) e CSI em função da classificação das unidades geomecânicas.

6.2. Dados sísmicos (Ocean Bottom Nodes)

Os dados sísmicos utilizados neste trabalho consistem em volumes de de impedância da onda P (Zp), impedância da onda S (Zs), razão Vp/Vs, Carbonate-Shale Indicator (CSI) e densidade volumétrica da rocha (RHOB) obtidos de inversão sísmica realizada pela equipe do Laboratório de Geofísica da Unicamp. Os dados utilizados para a inversão sísmica consistiram em angle-stacks de levantamentos de nós do fundo do oceano (OBN).

Os levantamentos sísmicos tridimensionais são realizados visando o imageamento das estruturas geológicas de subsuperfície (Kearey *et al*., 2002). Nesse sentido, é organizado um arranjo de tiros e receptores de forma a permitir a reunião de grupos de chegadas registradas que representam conjuntos de raios refletidos por uma determinada área de cada interface refletora (Kearey *et al*., 2002).

Nos sistemas convencionais de levantamentos sísmicos 3D, os sensores sísmicos são posicionados na superfície terrestre ou em plataformas marítimas. Por outro lado, os nós do fundo do oceano (OBN) são alocados diretamente no leito marinho. Nesse sentido, os dados de OBN apresentam uma resolução sísmica melhor que os dados de levantamentos convencionais devido ao levantamento de OBN eliminar ruídos e distorções introduzidos pela interface entre ar e água (Detomo *et al*., 2012).

6.3. Softwares

A aplicação do algoritmo de classificação foi feita através de programação em linguagem Python por meio das workstations disponíveis no Laboratório de Geofísica do IG/UNICAMP. A Tabela 3 apresenta a configuração da workstation utilizada neste trabalho. Os resultados de classificação foram carregados no software Jason Workbench (Geosoftware®) permitindo a visualização 3D dos modelos gerados.

Tabela 3: Resumo das configurações da estação de trabalho utilizada nesta pesquisa.

Sistema Operacional	Windows 11 Pro for Workstations
Tipo de Sistema	64 bits
Processador	Intel(R) Xeon(R) W-2145 CPU @ 3.70GHz
Memória RAM	128 GB
Núcleos	8
Processadores lógicos	16
Capacidade de Armazenamento	2,3 TB

7. METODOLOGIA

A Figura 14 apresenta o fluxograma aplicado neste trabalho. Inicialmente, os dados obtidos de inversões sísmicas foram utilizados para extração de novas variáveis. A inversão sísmica foi realizada previamente pela equipe do Laboratório de Geofísica da Unicamp a partir dos dados angle stack do conjunto OBN.

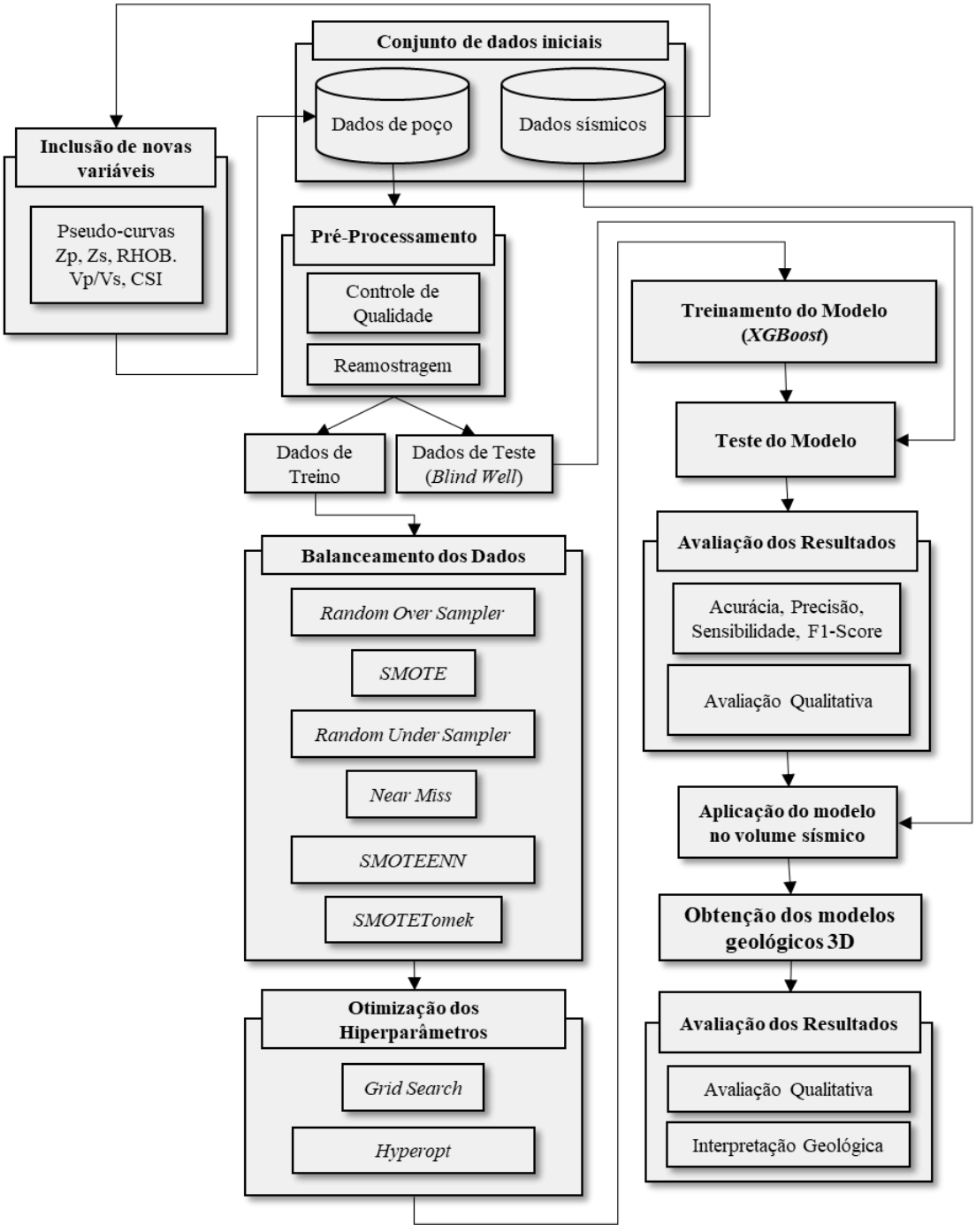

Figura 14: Fluxograma aplicado neste trabalho para a obtenção de modelos 3D.

Os resultados da inversão sísmica (Zp, Zs, RHOB, VP/VS, CSI) foram utilizados para extrair pseudocurvas dessas variáveis para os poços estudados nesta pesquisa. Com isso, além dos dados dessas variáveis na escala de poço, também contamos com as curvas na escala sísmica.

Para a classificação foi utilizado o algoritmo Extreme Gradient Boosting (XGBoost) proposto por Chen e Guestrin (2016). O algoritmo foi aplicado para classificar 3 alvos: Fácies Petroelásticas, Unidades de Fluxo e Fácies Geomecânicas. Para cada variável foi selecionado um poço para *blind test* dentre os disponíveis (Tabela 1) até que todos os poços tenham sido utilizados para o teste.

A classificação das Unidades de Fluxo (UF) conta com uma maior limitação na quantidade de dados disponíveis, havendo apenas 5 poços para treinamento. Desses 5 poços, apenas 3 apresentam as 4 unidades de fluxo, em proporção abaixo de 1%. Por conta dessa restrição, para a classificação das UF, foi adotado outro método para separação de dados de treinamento e de teste. Neste modelo, os dados de treinamento correspondem a 75% dos dados totais, utilizando todos os poços, enquanto os dados de teste correspondem aos 25% restantes.

Uma vez que o treinamento de modelos baseados em aprendizagem de máquina usando um banco de dados sub-representativo e criticamente não-balanceado pode limitar o desempenho do algoritmo (He & Garcia, 2009), avaliou-se o desempenho do modelo ao ser treinado com dados previamente balanceados por diferentes técnicas de reamostragem.

Foram aplicados dois algoritmos de sobreamostragem (*Random Over Sampler* e SMOTE), dois algoritmos de subamostragem (*Random Under Sampler* e *Near Miss*) e dois algoritmos híbridos (SMOTTomek e SMOTEENN) a fim de balancear os dados de treino de cada combinação de alvo e poço de teste. Adicionalmente foi avaliado o desempenho do modelo sem o balanceamento prévio das classes. Para a classificação de unidades de fluxo, devido à limitação das amostragem da UF4, foram aplicados apenas os algoritmos baseados em sobreamostragem e híbridos.

O princípio dos algoritmos de sobreamostragem de classes é o aumento artificial de amostras das classes menos amostradas. O algoritmo *Random Oversampler* (Japkowicz, 2000) funciona por meio da duplicação de amostras das classes minoritárias, selecionadas aleatoriamente, até

que as mesmas atinjam um equilíbrio com as classes mais amostradas. O SMOTE (*Synthetic Minority Over-sampling Technique*) (Chawla et al., 2002) utiliza amostras das classes menos amostradas em conjunto com amostras vizinhas da mesma classe para criar dados sintéticos por meio da interpolação do amostra original com dados das amostras vizinhas, levando em consideração a proximidade entre as amostras, até que o equilíbrio seja atingido.

Por outro lado, os algoritmos de subamostragem se baseiam na redução de amostras das classes com maior quantidade de dados. O *Random Undersampler* (Japkowicz, 2000) seleciona aleatoriamente dados das classes mais amostradas para serem mantidos em conjunto de dados reduzidos, até que este novo conjunto esteja em equilíbrio com as classes menos amostradas. O *Near Miss* (Mani & Zhang, 2003) calcula a distância entre amostras das classes majoritárias e minoritárias em relação a alguma característica das amostras e mantém cada amostra da classe majoritária que esteja mais próxima da classe minoritária em termos da característica escolhida, criando um novo conjunto de dados reduzido.

Em alternativa, os algoritmos híbridos combinam técnicas de sobreamostragem e de subamostragem. Nesse sentido, o algoritmo SMOTEENN (Chawla et al., 2002) combina as técnicas do SMOTE e do Edited Nearest Neighbors para lidar com conjuntos de dados desbalanceados. Primeiro, o SMOTE é aplicado, sintetizando novas amostras para as classes minoritárias. Em seguida, o Edited Nearest Neighbors é utilizado para remover amostras das classes majoritárias, resultando em um conjunto de dados mais equilibrado e reduzido.

A última técnica de balanceamento avaliada neste trabalho consiste no algoritmo SMOTETomek (Wang et al., 2019), que combina as técnicas do SMOTE e do Tomek Links para remover exemplos ruidosos das classes majoritárias e sintetizar novas amostras para as classes minoritárias. Primeiro, o SMOTE é aplicado para criar novas amostras para as classes minoritárias. Em seguida, o método Tomek Links é utilizado para identificar e remover pares de exemplos de classes majoritárias e minoritárias que estão muito próximos entre si, visando aprimorar a separabilidade entre as classes.

Após os ajustes dos dados de treinamento, a biblioteca hyperopt foi aplicada para a otimização dos hiperparâmetros do modelo. Após a avaliação dos resultados da classificação por meio do conjunto de dados de teste, o modelo foi aplicado no volume de dados obtidos da inversão sísmica. Com isso, obteve-se modelos tridimensionais de distribuição de fácies petroelásticas, unidades de fluxo e unidades geomecânicas.

A partir da obtenção dos resultados, a função plot_importance do XGBoost foi aplicada para o visualizar o peso de cada variável na classificação. O XGBoost utiliza três parâmetros para calcular a importância das variáveis: peso (*weight*), ganho (*gain*) e cobertura (*cover*) (XGBoost Doc., 2022).

O peso consiste no número de vezes em que a variável aparece nas árvores. O ganho representa a melhoria do desempenho de cada variável ao ser aplicada para fazer uma nova divisão, sendo calculado pelo índice de impureza de Gini. A cobertura corresponde à quantidade de vezes que a variável é usada para fazer novas divisões nas árvores. A importância final de cada variável é calculada pelo produto do ganho médio da variável pelo valor normalizado de sua cobertura (XGBoost Doc., 2022).

7.1. Avaliação dos resultados

Modelos de classificação são frequentemente avaliados por meio de matrizes de confusão (Figura 15), que apresentam os resultados da classificação em quatro categorias (Kubat, 2017):

- Verdadeiro Positivo (VP): Quando o modelo classifica corretamente o dado como positivo.
- Verdadeiro Negativo (VN): Quando o modelo classifica corretamente o dado como negativo.
- Falso Positivo (FP): Quando o modelo classifica o dado como positivo mas na verdade o dado é negativo.
- Falso Negativo (FN): Quando o modelo classifica o dado como negativo mas na verdade o dado é positivo.

		Valor Predito	
		Positivo	Negativo
Valor Real	Positivo	VP	FN
	Negativo	FP	VN

Figura 15: Exemplo de matriz confusão.

Com base nessas categorias são determinados os valores de Acurácia (Equação 7.1), Precisão (Equação 7.2), Sensibilidade (Equação 7.3) e F1-Score (Equação 7.4), métricas comumente utilizadas para avaliação de modelos de classificação (Kubat. 2017).

A acurácia apresenta a proporção de classificações corretas dentre o total de amostras, fornecendo uma visão geral do resultado da classificação.

$$\text{Acurácia} = \frac{VP + VN}{VP + VN + FP + FN} \qquad (7.1)$$

A precisão indica a capacidade do modelo de identificar corretamente os casos positivos, dentre o total de casos classificados como positivos. Dessa forma, a precisão permite avaliar a confiabilidade das predições de cada classe no modelo, uma vez que indica a proporção de falso-negativo. Nesse sentido, a precisão das classes que indicam bons reservatórios deve ser muito alta, visando reduzir a probabilidade de perfuração de poços secos ou marginais.

$$\text{Precisão} = \frac{VP}{VP + FP} \qquad (7.2)$$

A sensibilidade indica a capacidade do modelo de identificar corretamente os casos positivos dentre a quantidade real de casos positivos, importante para avaliar o desempenho em identificar todas as reservas disponíveis.

$$\text{Sensibilidade} = \frac{VP}{VP + FN} \qquad (7.3)$$

O F1-Score é uma média harmônica entre a precisão e a sensibilidade, sendo importante para identificar limitações do classificador quando há dados não-balanceados.

$$\text{F1-Score} = \frac{2 \times Precisão \times sensibilidade}{Precisão + sensibilidade} \qquad (7.4)$$

8. RESULTADOS

As Figura 16, 17 e 18 apresentam o resumo dos resultados de Acurácia, Precisão, Sensibilidade (*Recall*) e F1-Score para cada alvo, agrupados por poços de teste e por algoritmos de reamostragem. Esses valores correspondem à média aritmética dos resultados obtidos utilizando cada combinação de poço de teste e algoritmo de reamostragem.

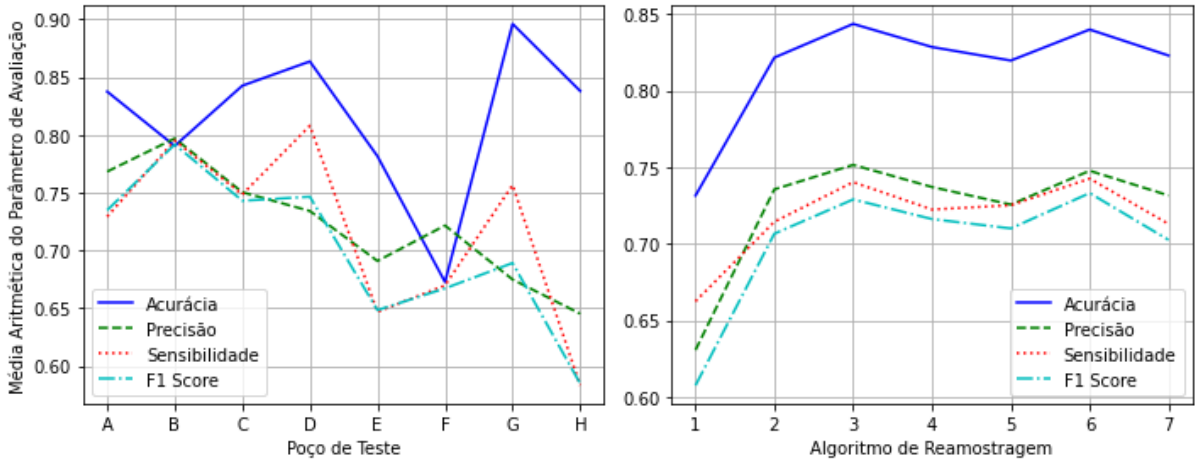

Figura 16: Resumo dos resultados de classificação de Fácies Petroelásticas agrupados por poço de teste (à esquerda) e por algoritmo de reamostragem (à direita). Algoritmos de Reamostragem: **1** - Near Miss; **2** - Random Over Sampler; **3** - Random Under Sampler; **4** - SMOTE; **5** - SMOTEEN; **6** - SMOTETomek; **7** - Sem a aplicação de algoritmo de reamostragem.

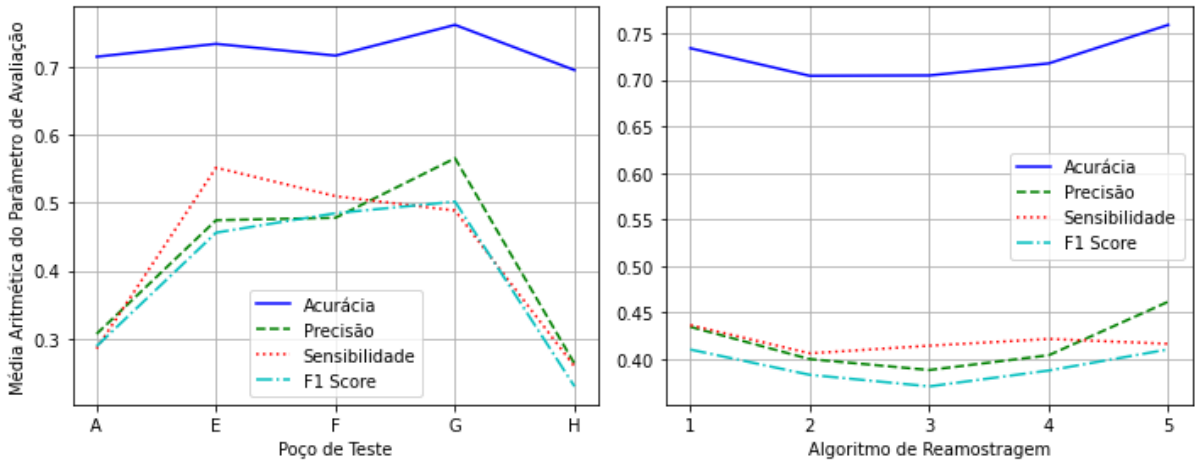

Figura 17: Resumo dos resultados de classificação de Unidades de Fuxo agrupados por poço de teste (à esquerda) e por algoritmo de reamostragem (à direita). Algoritmos de Reamostragem: **1** - Random Over Sampler; **2** - SMOTE; **3** - SMOTEEN; **4** - SMOTETomek; **5** - Sem a aplicação de algoritmo de reamostragem.

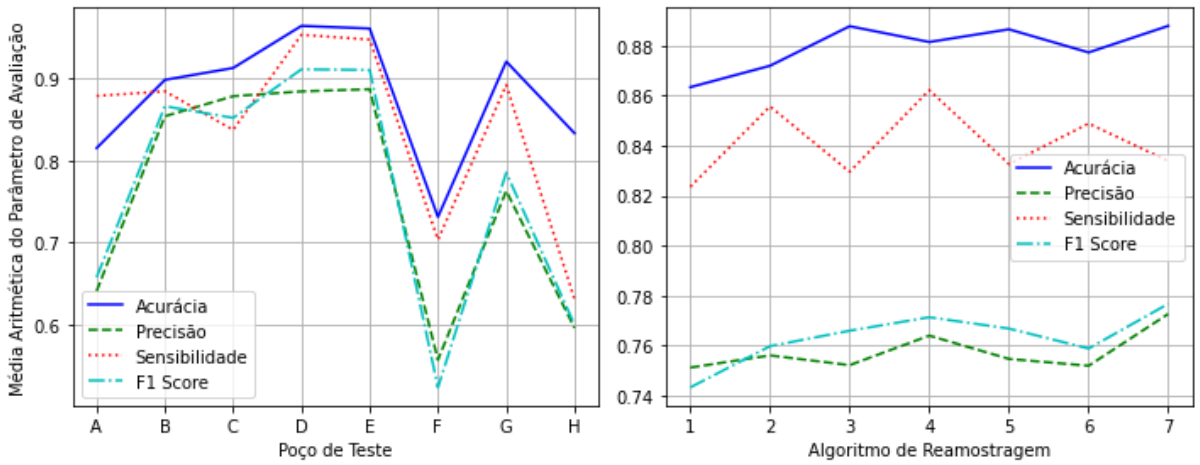

Figura 18: Resumo dos resultados de classificação de Unidades Geomecânicas agrupados por poço de teste (à esquerda) e por algoritmo de reamostragem (à direita). Algoritmos de Reamostragem: **1** - Near Miss; **2** - Random Over Sampler; **3** - Random Under Sampler; **4** - SMOTE; **5** - SMOTEEN; **6** - SMOTETomek; **7** - Sem a aplicação de algoritmo de reamostragem.

Por se tratar de dados não balanceados, as médias gerais de avaliação de desempenho do modelo podem mascarar limitações do modelo em classificar as classes menos amostradas. Nesse sentido, as matrizes de confusão apresentam os resultados para cada classe, permitindo uma avaliação mais detalhada. As figuras 19, 20 e 21 apresentam as matrizes de confusão das classificações de fácies petroelásticas, unidades de fluxo e unidades geomecânicas, respectivamente.

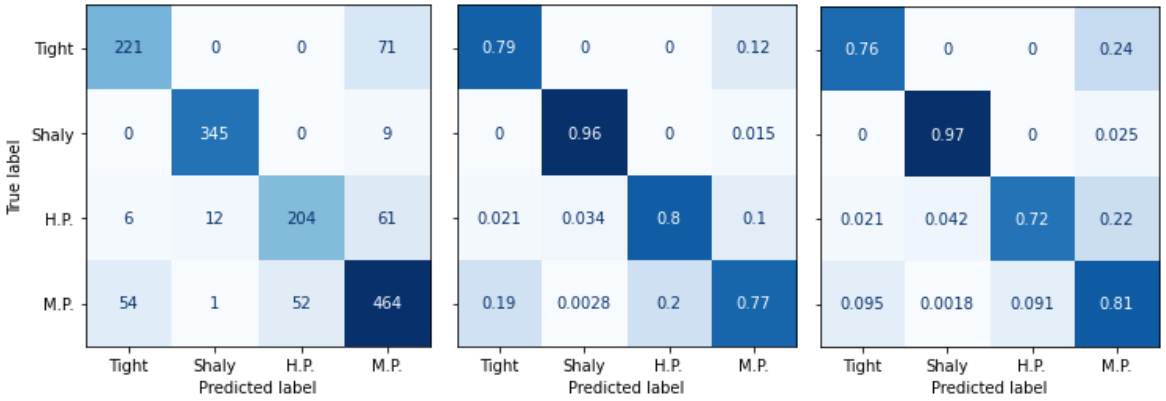

Figura 19: Matrizes de confusão da classificação de fácies petroelásticas. Da esquerda para direita: Matriz sem normalização; Normalizada pelos valores preditos (Precisão); Normalizada pelos valores verdadeiros (Sensibilidade).

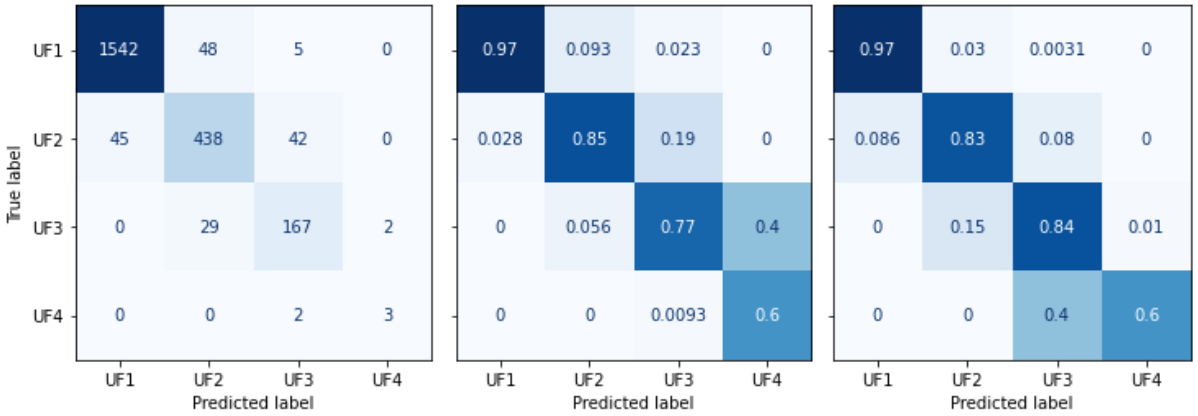

Figura 20: Matrizes de confusão da classificação de unidades de fluxo. Da esquerda para direita: Matriz sem normalização; Normalizada pelos valores preditos (Precisão); Normalizada pelos valores verdadeiros (Sensibilidade).

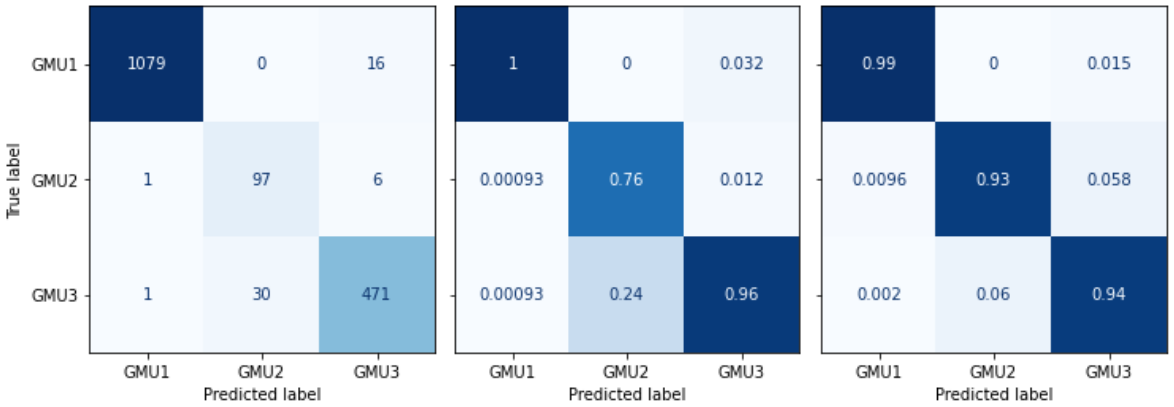

Figura 21: Matrizes de confusão da classificação de unidades geomecânicas. Da esquerda para direita: Matriz sem normalização; Normalizada pelos valores preditos (Precisão); Normalizada pelos valores verdadeiros (Sensibilidade).

As Figuras 22, 23 e 24 apresentam seções exibindo os resultados da classificação na escala sísmica dos modelos de fácies petroelásticas, unidades de fluxo e unidades geomecânicas, assim como os dados de classificação prévia dos poços, mostrando uma correlação significativa entre o resultado da classificação e os dados de poços.

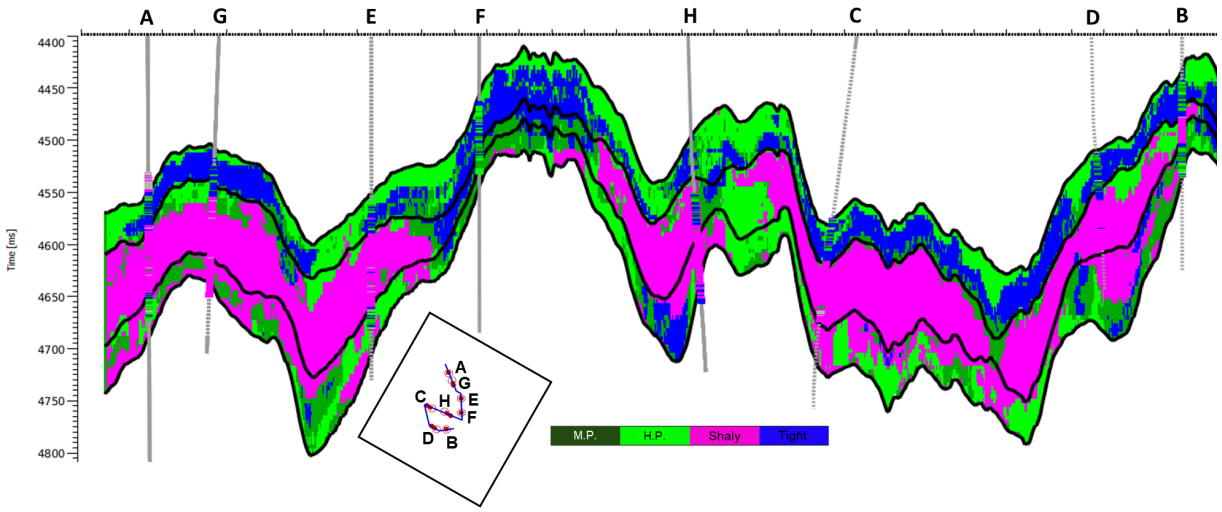

Figura 22: Fácies petroelásticas classificadas em uma seção do volume sísmico. O Poço B corresponde ao *Blind Test* deste modelo. Esta seção é composta por vários segmentos 2D extraídos do volume 3D ao longo de diferentes direções para cruzar os oito poços, tendo uma extensão total de 9 km. O mapa abaixo da seção exibe o caminho representado pela seção, além da posição dos poços.

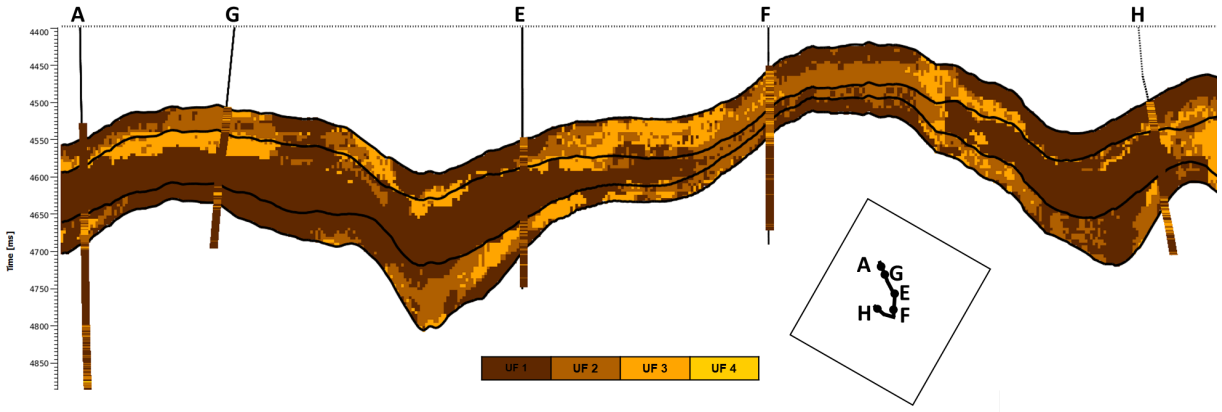

Figura 23: Unidades de Fluxo classificadas em uma seção do volume sísmico. Para esta classificação, foram utilizados 75% dos dados de poços para treinamento, e 25% para teste.

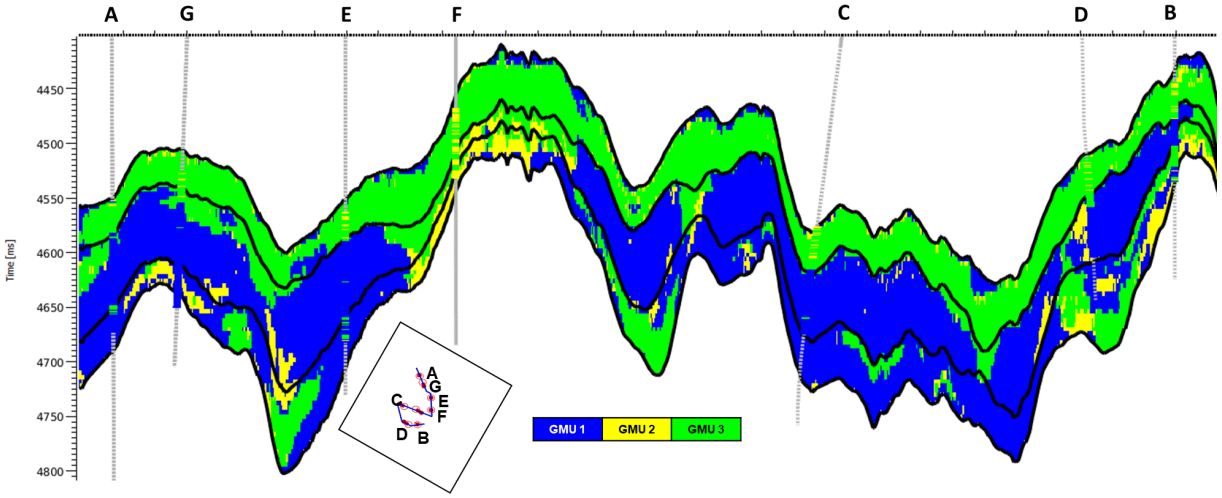

Figura 24: Unidades Geomecânicas classificadas em uma seção do volume sísmico. O Poço B foi utilizado para *Blind Test* neste modelo.

As Figuras 25, 26 e 27 apresentam seções de crossline do modelo, exibindo uma visão geral da distribuição das classes em 3D.

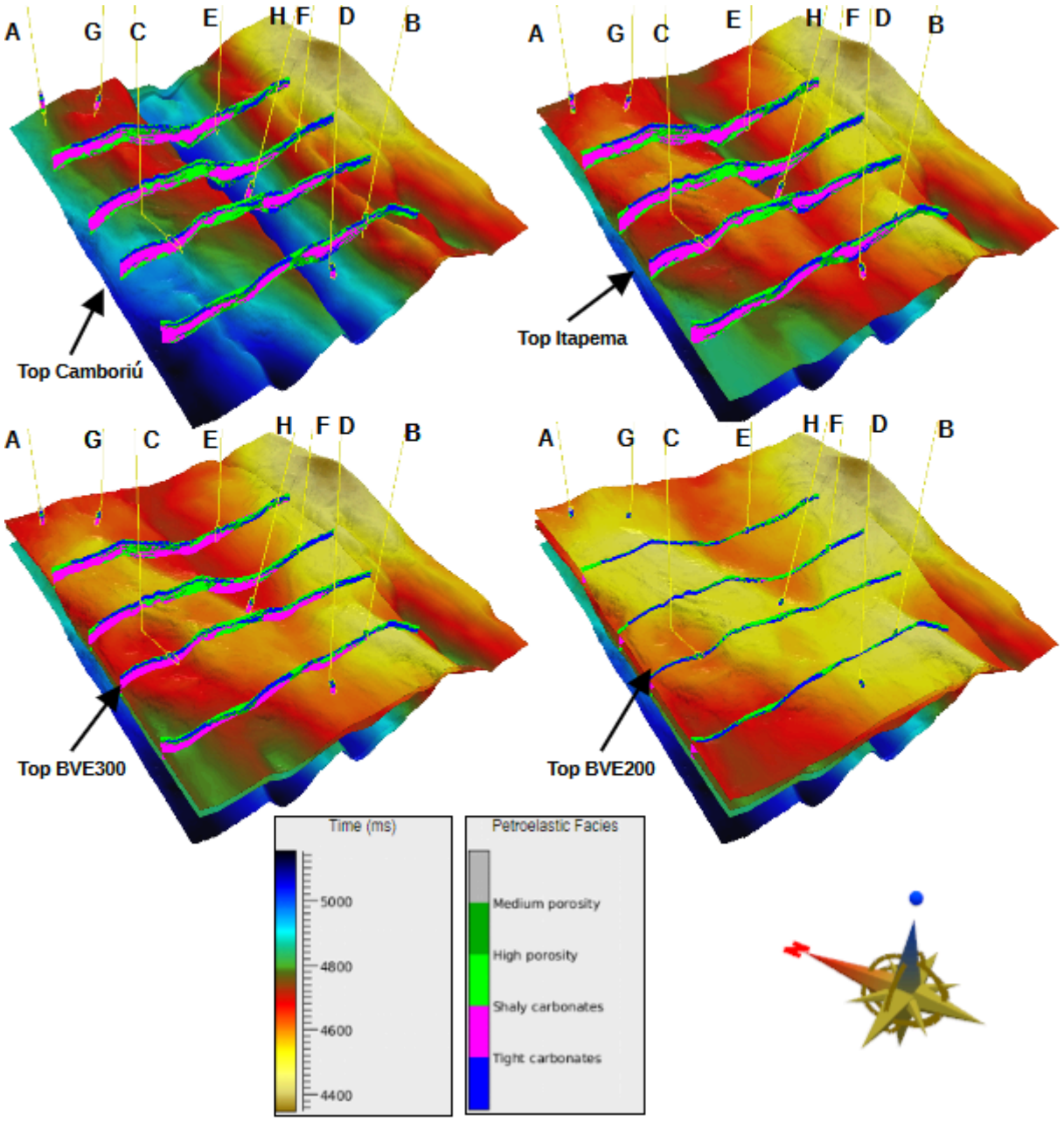

Figura 25: Fácies Petroelásticas classificadas em seções do volume sísmico, poços e horizontes do topo das formações Camboriú e Itapema e dos intervalos de produção BVE 100 e BVE 200.

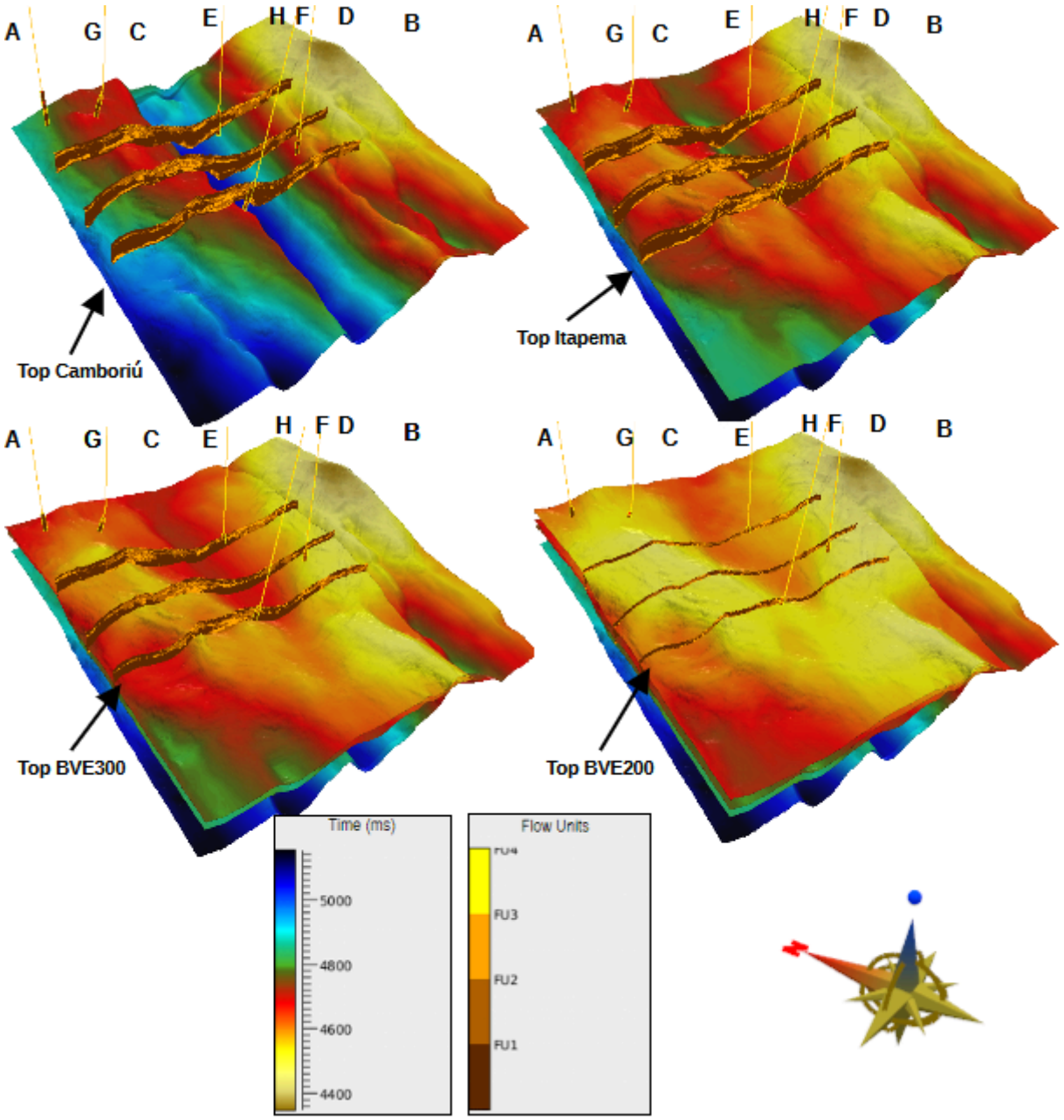

Figura 26: Unidades de Fluxo classificadas em seções do volume sísmico, poços e horizontes do topo das formações Camboriú e Itapema e dos intervalos de produção BVE 100 e BVE 200.

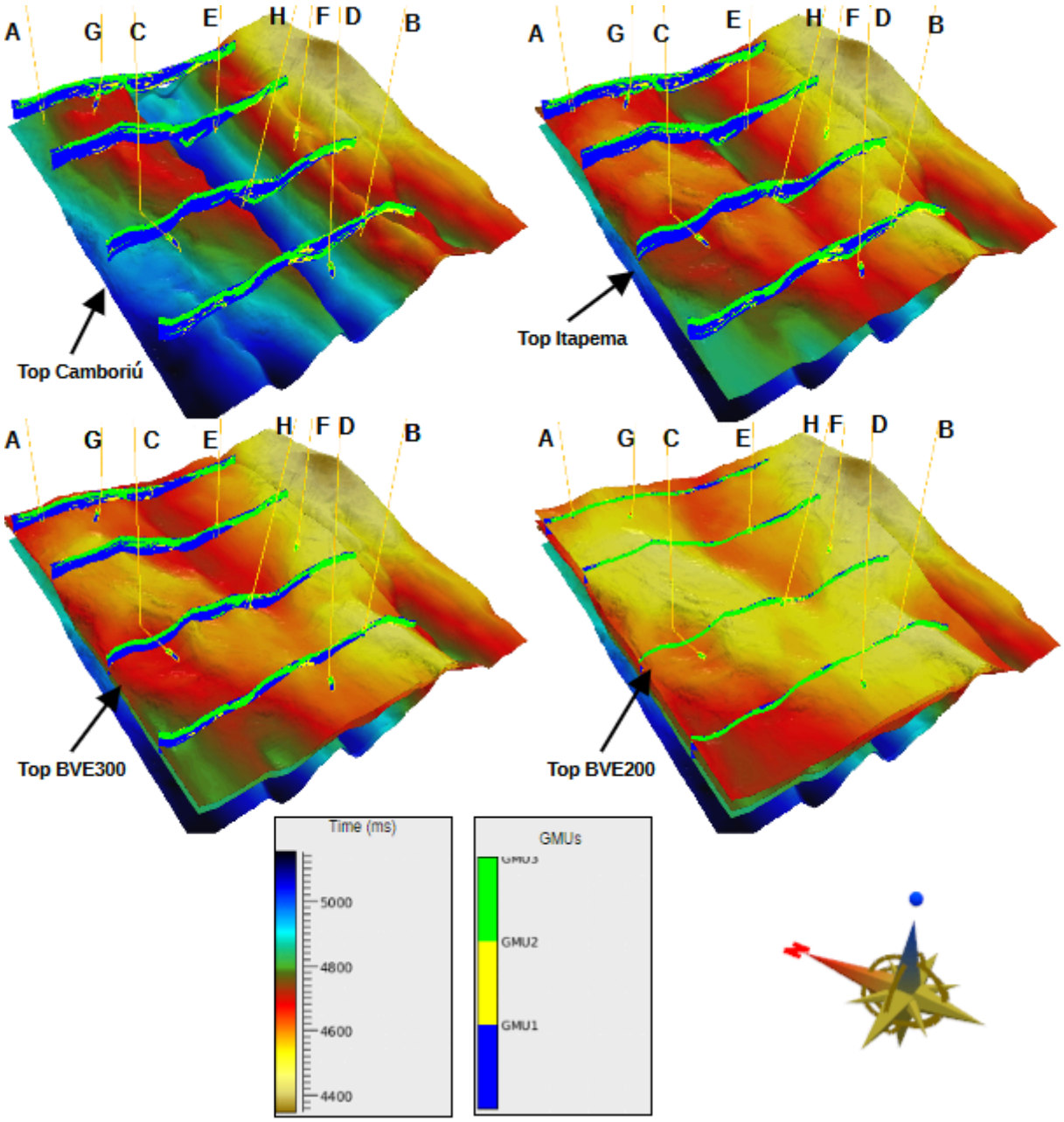

Figura 27: Unidades Geomecânicas classificadas em seções do volume sísmico, poços e horizontes do topo das formações Camboriú e Itapema e dos intervalos de produção BVE 100 e BVE 200.

A Figura 28 apresenta os resultados de classificação e diferentes produtos de inversão sísmica em uma seção do modelo, exibindo uma visão geral da distribuição das classes.

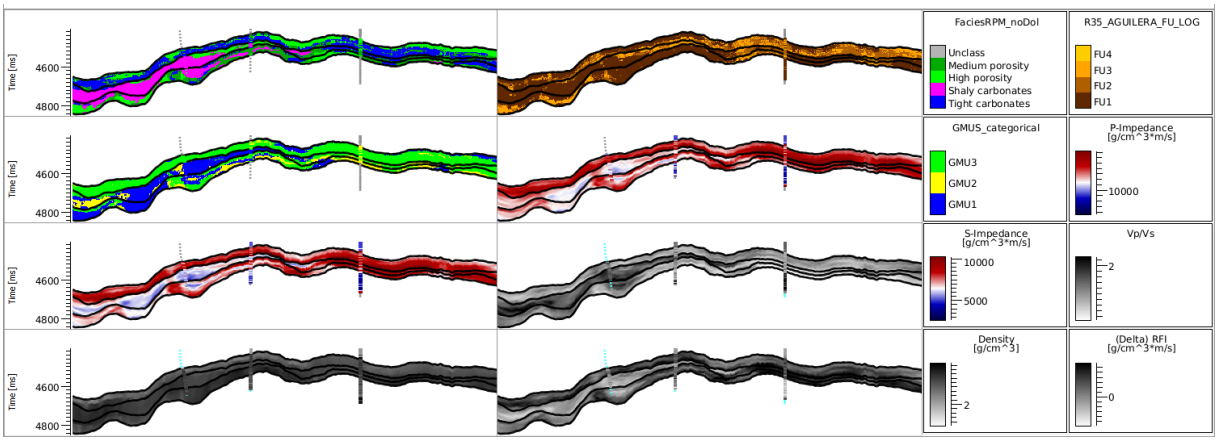

Figura 28: Resultados das classificações com os poços B, D e F (da esquerda para a direita) e os produtos de inversão sísmica utilizados neste trabalho para geração dos modelos 3D

As Figuras 29, 30 e 31 apresentam diagramas de dispersão dos parâmetros elásticos, previamente obtidos por inversão sísmica, destacando os resultados de classificação. Os resultados da classificação de fácies petroelásticas estão coerentes com os diagramas de dispersão dos dados de poços apresentados na Figura 11.

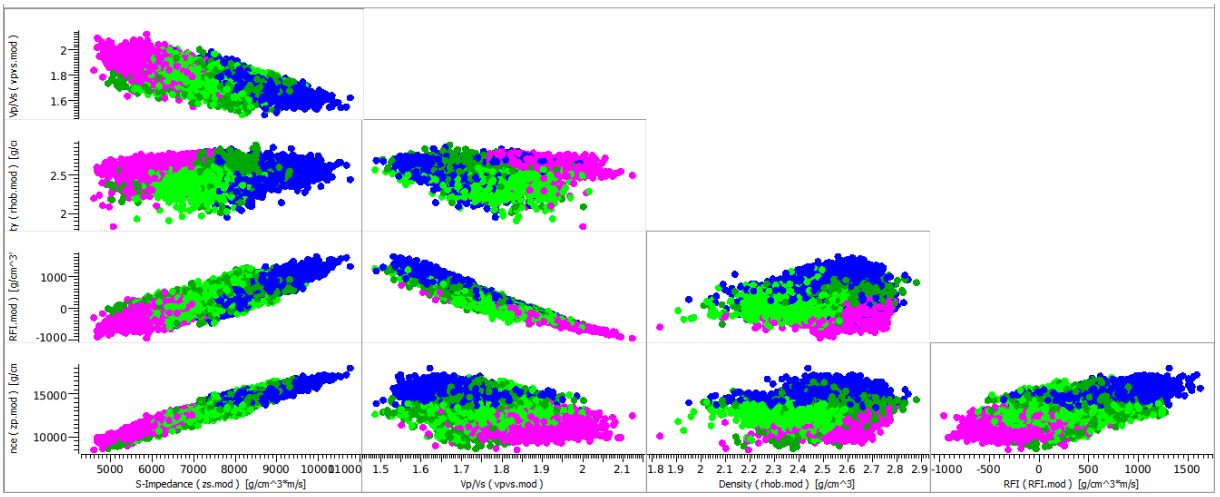

Figura 29: Diagramas de dispersão do resultado da classificação de Fácies Petroelásticas no volume de dados sísmicos. Variáveis do eixo Y, de baixo para cima: Zp, CSI, RHOB e Vp/Vs. Variáveis do eixo X, da esquerda para direita: Zs, Vp/Vs, RHOB e CSI.

Por outro lado, na classificação de Unidades de Fluxo, verifica-se um grande número de amostras de baixa densidade e baixo Vp/Vs classificadas como Unidade de Fluxo

1, o que não é identificado nos dados de poços (Figura 12). Nos dados de poços, os menores valores de densidade e Vp/Vs são apresentados pelas UFs 2, 3 e 4. Portanto, é verificada uma limitação do modelo em classificar com precisão as Unidades de Fluxo da Formação Barra Velha.

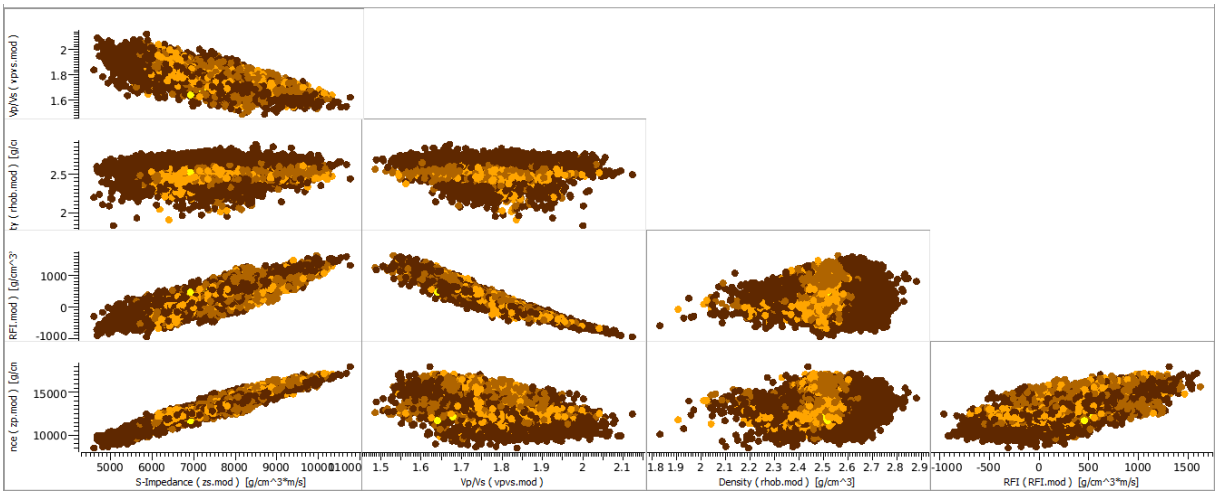

Figura 30: Diagramas de dispersão do resultado da classificação de Unidades de Fluxo no volume de dados sísmicos. Variáveis do eixo Y, de baixo para cima: Zp, CSI, RHOB e Vp/Vs. Variáveis do eixo X, da esquerda para direita: Zs, Vp/Vs, RHOB e CSI.

Os resultados obtidos com a classificação das unidades geomecânicas condizem com os diagramas de dispersão dos dados de poço (Figura 13), apontando o potencial do modelo para esse tipo de classificação.

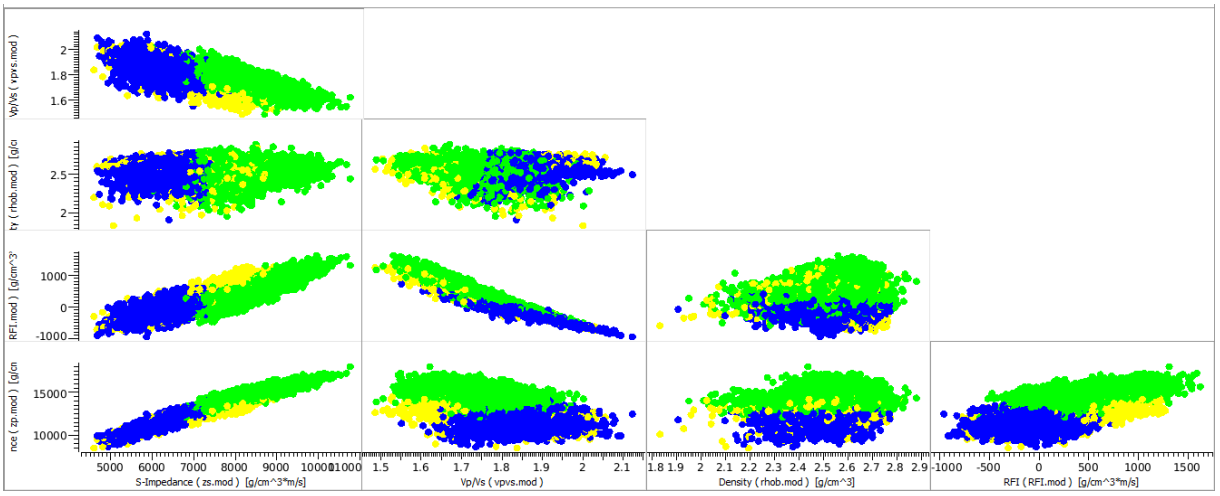

Figura 31: Diagramas de dispersão do resultado da classificação de Unidades Geomecânicas no volume de dados sísmicos. Variáveis do eixo Y, de baixo para cima: Zp, CSI, RHOB e Vp/Vs. Variáveis do eixo X, da esquerda para direita: Zs, Vp/Vs, RHOB e CSI.

A tabela 4 apresenta a média geral dos resultados de Acurácia, Precisão, Sensibilidade e F1-Score em dois cenários. O primeiro consiste no treinamento do modelo utilizando as variáveis Zp, Zs, RHOB, Vp/Vs e CSI no treinamento. Portanto, utiliza apenas dados em escala de poço no treino. O segundo cenário consiste no treinamento do modelo utilizando as variáveis anteriores em conjunto com as pseudocurvas extraídas dos dados sísmicos, resultando no conjunto: Zp, Zs, RHOB, Vp/Vs, CSI, Zp_pc, Zs_pc, RHOB_pc, Vp/Vs_pc, e CSI_pc. Dessa forma, este cenário combina dados em escala de poço e escala sísmica durante o treinamento do modelo.

Tabela 4: Resumo dos resultados de classificação utilizando i) as curvas tradicionais e ii) as curvas tradicionais em conjunto com as pseudocurvas.

Variáveis utilizadas no treinamento do modelo	Acurácia Média	Precisão Média	Sensibilidade Média	F1-Score Médio
Zp, Zs, RHOB, Vp/Vs e CSI	0.81	0.68	0.72	0.67
Zp, Zs, RHOB, Vp/Vs, CSI, Zp_pc, Zs_pc, RHOB_pc, Vp/Vs_pc, e CSI_pc	0.82	0.68	0.71	0.67

De forma geral, verifica-se que a utilização das pseudocurvas no treinamento do modelo não resultou em uma melhoria expressiva das métricas de avaliação de desempenho dos modelos.

Adicionalmente foi avaliado o impacto das pseudocurvas na distribuição das amostras dos resultados de classificação de cada intervalo de produção (Figura 32). Os poços da Formação Barra Velha geralmente apresentam três intervalos de produção: Barra Velha 100 (BVE100), Barra Velha 200 (BVE200) e Barra Velha 300 (BVE300) (Rebelo et al., 2022). O BVE100 é caracterizado por altas porosidades e permeabilidades, consistindo em um excelente reservatório (Rebelo et al., 2022). O BVE200 consiste em uma zona microporosa, com permeabilidades extremamente baixas (Rebelo et al., 2022). O BVE300 geralmente é um tipo de reservatório mais pobre (Rebelo et al., 2022).

Os gráficos permitem identificar que os modelos que foram treinados com pseudocurvas resultaram em distribuições mais realistas uma vez que as unidades que indicam

rochas reservatórios apresentam menor proporção nos resultados gerados por este treinamento.

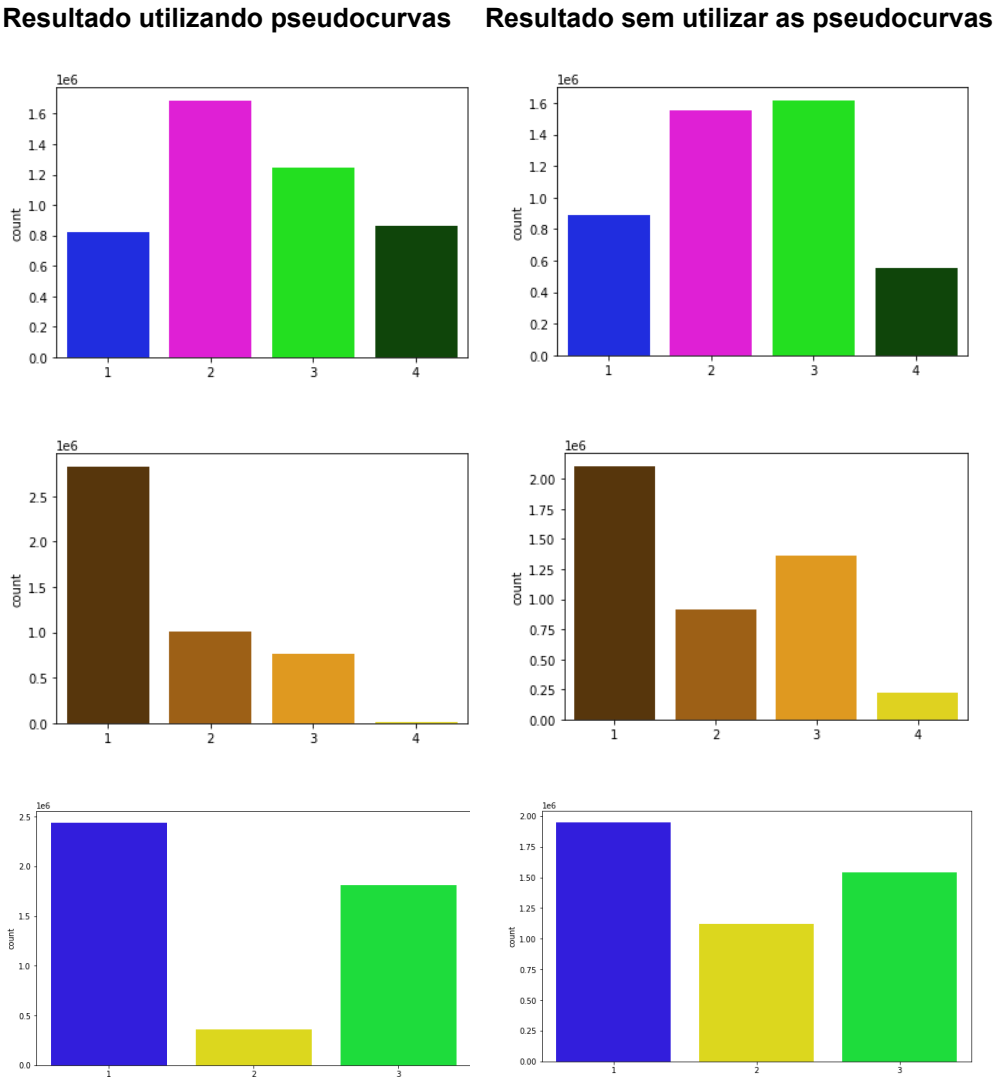

Figura 32: Histogramas apresentando a distribuição das fácies petroelásticas (primeira linha), unidades de fluxo (segunda linha) e unidades geomecânicas (terceira linha). A primeira coluna corresponde à distribuição das classes no resultado da aplicação do modelo no volume sísmico, utilizando as pseudocurvas (Zp_pc, Zs_pc, RHOB_pc, VpVs_pc e CSI_pc) no treinamento do modelo. A terceira coluna corresponde à distribuição das classes no resultado da aplicação do modelo sem utilizar as pseudocurvas no treinamento.

A Figura 33 apresenta os histogramas de distribuição das amostras de cada classificação para cada intervalo de produção da Formação Barra Velha: BVE100, BVE200 e BVE300.

Com relação aos modelos 3D obtidos pela classificação dos produtos de inversão sísmica, verificou-se nas fácies petroelásticas, a predominância das fácies Tight e High Porosity no BVE100. No BVE200, a predominância é da fácies Shaly. No BVE 300, das fácies High Porosity, Medium Porosity e Shaly, nesta ordem, com pequena variação entre as proporções.

Com relação à classificação de unidades de fluxo, a UF1 predomina nos três intervalos de produção da Formação Barra Velha. Nos intervalos BVE100 e BVE300 as proporções das unidades U2 e UF3 são maiores que no BVE200, sendo que no BVE100 é onde ocorre a maior proporção de UF3.

Na classificação de unidades geomecânicas, a unidade GMU3 predomina no BVE100, a GMU1 predomina no BVE200, e no BVE300 ocorre, de forma aproximada, um equilíbrio entre GMU1 e GMU3, sendo que a GMU1 apresenta uma proporção maior.

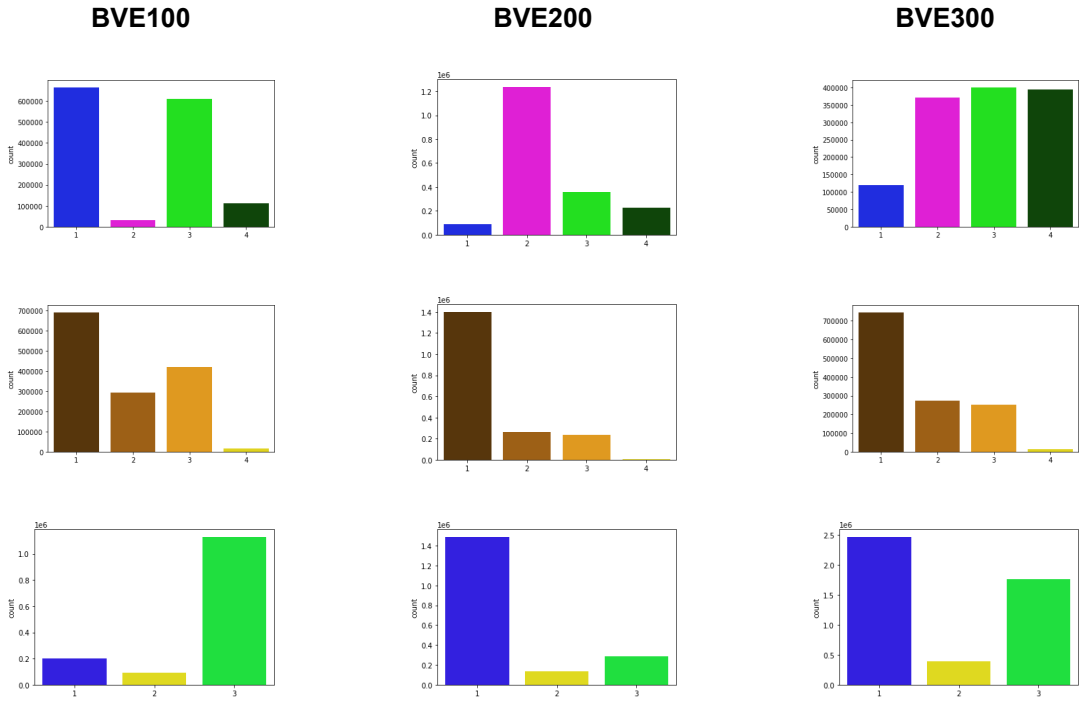

Figura 33: Histogramas apresentando a distribuição dos resultados de classificação de fácies petroelásticas (primeira linha), unidades de fluxo (segunda linha) e unidades geomecânicas (terceira linha), para cada intervalo de produção da Formação Barra Velha: BVE100 (primeira coluna), BVE200 (segunda coluna) e BVE300 (terceira coluna).

A função plot_importance do XGBoost foi aplicada a fim de avaliar o peso de cada variável na classificação de cada modelo (Figura 34). Os gráficos abaixo apresentam o ganho de cada variável. O ganho representa a melhoria do desempenho ao aplicar a variável para fazer uma nova divisão (XGBoost Doc., 2022).

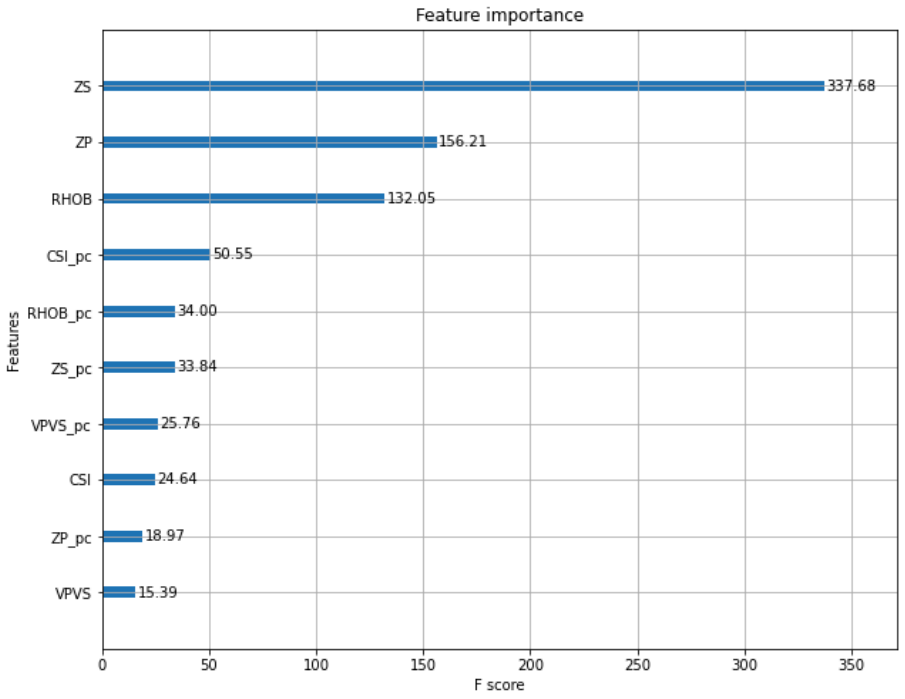

Figura 34: Importância de cada atributo para a classificação de fácies petroelásticas.

Para os dados de unidade de fluxo, a aplicação plot_importance indicou os dados de pseudocurvas de Zs como o maior ganho para a classificação, seguido da pseudo-curva de carbonate-shale indicator (CSI) (Figura 35).

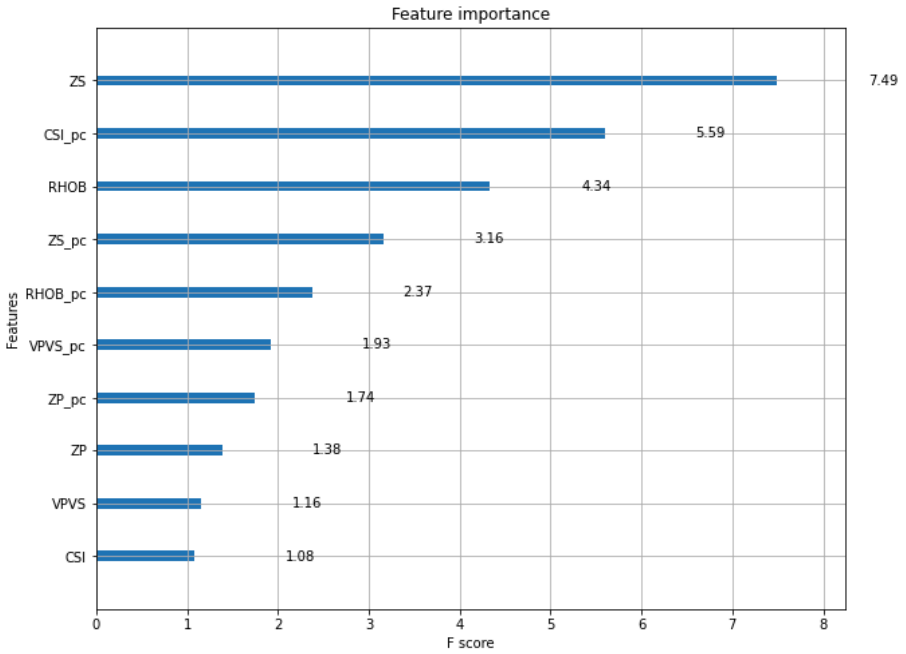

Figura 35: Importância de cada atributo para a classificação de unidades de fluxo

A aplicação do plot_importance no modelo de classificação das unidades geomecânicas indica uma grande importância das pseudocurvas na classificação. Após as impedâncias P e S, classificadas com maior ganho, as variáveis seguintes correspondem às pseudocurvas (Figura 36).

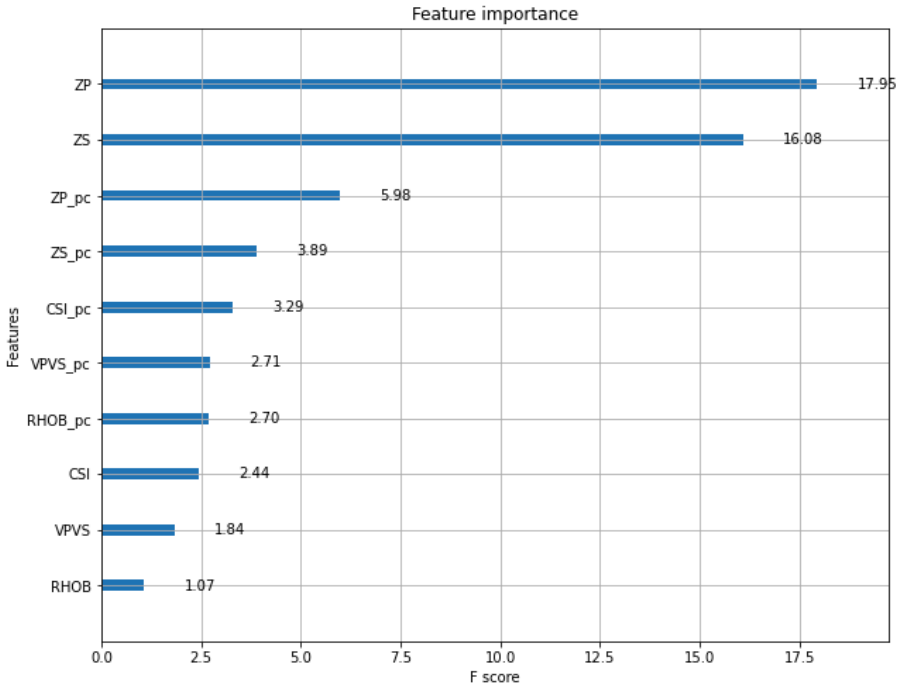

Figura 36: Importância de cada atributo para a classificação de unidades geomecânicas.

9. DISCUSSÃO

De modo geral, a acurácia dos modelos foi superior a 70% na maioria dos conjuntos de alvo, poço de teste e algoritmo de classificação. Entretanto, a avaliação somente pela acurácia pode induzir a uma avaliação incorreta do desempenho do classificador, uma vez que a classificação de dados não-balanceados pode resultar em uma acurácia geral alta ainda que a classe minoritária não seja classificada corretamente. Isso foi observado neste trabalho na classificação das unidades de fluxo, que teve uma acurácia geral acima de 70% entretanto somente 60% da classe UF 4 foi classificada corretamente. Por este motivo, os resultados das classificações também foram avaliados por meio das métricas acurácia, precisão, sensibilidade e F1-Score.

A acurácia das classificações de fácies petroelásticas está na faixa de 67 a 90%. Os melhores resultados foram encontrados para os modelos que utilizaram como *blind test* os poços A, B, C e D, nos quais todas as métricas estão na faixa de 73 a 87%. Os poços E, F e G apresentaram resultados superiores a 64% para todas as métricas. Enquanto o poço H teve o pior desempenho, apesar da acurácia próxima a 85%, as métricas de precisão, sensibilidade e F1 score apresentaram valores abaixo de 65%.

Esses resultados demonstram uma sensibilidade do modelo às alterações dos conjuntos de dados de treinamento. Nesse sentido, entende-se que, neste conjunto de dados, o poço H é indispensável para a captura das relações entre os dados de entrada e saída. Por outro lado, os resultados demonstram que as relações entre entrada e saída aprendidas pelo modelo com os dados do poço B, no geral, são capturadas pelo restante do conjunto de dados. Dessa forma, a retirada desse poço para *blind test* não resulta em uma queda significativa de desempenho do modelo como é observado com o poço H.

Com relação aos métodos de reamostragem, para a classificação de fácies petroelásticas, os resultados mostram métricas superiores a 70% na maioria dos cenários, incluindo o cenário sem reamostragem de dados. No geral, a aplicação do balanceamento levou a uma pequena melhora dos resultados de classificação, chegando a aumentar em até 3% as médias de precisão, sensibilidade, f1-score e acurácia. A exceção é o método Near Miss, em que o desempenho do modelo caiu significativamente.

A precisão das classificações de unidades de fluxo está na faixa de 30 a 60%, a sensibilidade, de 30 a 55%, e F1-Score de 25 a 50%. Devido ao baixo desempenho do modelo de classificação de unidades de fluxo ao ser dividido em dados de treino e teste por meio da escolha de um poço de teste, foi adotada a divisão de dados de forma que 75% dos dados totais foram utilizados para treino e 25% para teste.

Entretanto, mesmo com essa abordagem, a classe UF 4 das Unidades de Fluxo, que indica o reservatório de maior qualidade, apresentou resultado de precisão e sensibilidade de 60%. Com isso, verifica-se uma limitação para a modelagem das unidades de fluxo. Contudo, a classe UF 4 apresenta a menor taxa de amostragem dentre as Unidades de Fluxo, correspondendo a menos de 1% do total de amostras, o que configura um alto desequilíbrio de classe (Leevy et al., 2018).

No geral, modelos baseados em aprendizagem de máquina apresentam melhor desempenho quando treinados com classes balanceadas. Entretanto, o resultado da classificação, após o treinamento com dados balanceados, depende criticamente do conjunto de dados disponíveis, sendo que a escolha da melhor abordagem dependerá de cada conjunto de dados (Wei & Dunbrack, 2013). No presente estudo, a aplicação dos algoritmos de balanceamento de dados não resultou em uma melhoria significativa dos resultados de classificação das unidades de fluxo.

No caso da classe UF4, por conta da baixa amostragem, o balanceamento por sobreamostragem é limitado uma vez que o algoritmo deve utilizar cerca de 30 amostras para geração de cerca de 4500 dados sintéticos. Além disso, o balanceamento por subamostragem também é dificultado uma vez que reduz drasticamente a quantidade total de dados de treino, resultando em uma piora geral do desempenho do modelo.

A precisão das classificações de unidades geomecânicas está na faixa de 55 a 90%, a sensibilidade, de 60 a 95%, e F1-Score de 55 a 92%. Os melhores resultados foram obtidos com os modelos que utilizaram como *Blind Test* os poços B, C, D e E, para os quais a média dos resultados de todas as métricas superam os 80%. Em seguida, os resultados com poço G como *blind test* apresentam médias entre 70 e 92%. Os poços A e H apresentam resultados entre 60 e 90%. O poço F apresenta as piores médias de resultados, com métricas abaixo de 60%.

Esses resultados mostram que neste conjunto de dados, os poços F, H e A são indispensáveis para o treinamento do modelo. Ao retirar esses poços do conjunto de dados de treino, o modelo não captura corretamente as relações entre as variáveis de entrada e saída, resultando na queda de desempenho.

Os resultados da aplicação de algoritmos de reamostragem na classificação de unidades de fluxo demonstram que para esse conjunto de dados, em média, a aplicação dos algoritmos de reamostragem não resultam em uma melhoria significativa de desempenho. Métricas como a sensibilidade, chegaram a subir 3% com a aplicação do SMOTE, entretanto com uma queda na precisão, f1-score e acurácia. Nesse contexto, deve-se levar em conta que a baixa precisão de uma classe favorável para exploração pode levar à perfuração de poços secos e marginais, enquanto a baixa sensibilidade pode resultar em reservatórios importantes não descobertos.

O resultado da análise de importância de variáveis indica que, na classificação de fácies petroelásticas, a maior importância é para a Impedância S, seguida da Impedância P. Além disso, uma parte significativa dos pesos das variáveis refere-se às pseudocurvas. Na classificação das unidades de fluxo, o principal ganho foi da pseudocurva de densidade. Na classificação de unidades geomecânicas, a maior importância foi de Zp, seguida das pseudocurvas de Zp e Zs. Outrossim, os histogramas de distribuição de classes (Figura 26) evidenciam a contribuição das pseudocurvas para a obtenção de resultados mais confiáveis na classificação de dados do volume sísmico. Nesse sentido, verifica-se a relevância desses atributos aplicados para melhoria do desempenho do classificador.

Mattos et al. (2022) demonstraram a correlação entre as fácies petroelásticas com as fácies litológicas interpretadas por Rebelo et al. (2022), com base da classificação de Gomes et al. (2020), em dados de lâminas e plugs da Formação Barra Velha (Figura 37). Mattos et al. (2022) identificaram que certas fácies petroelásticas apresentam correspondência direta com litofácies específicas. Isso sugere uma relação entre as propriedades petroelásticas e a litologia.

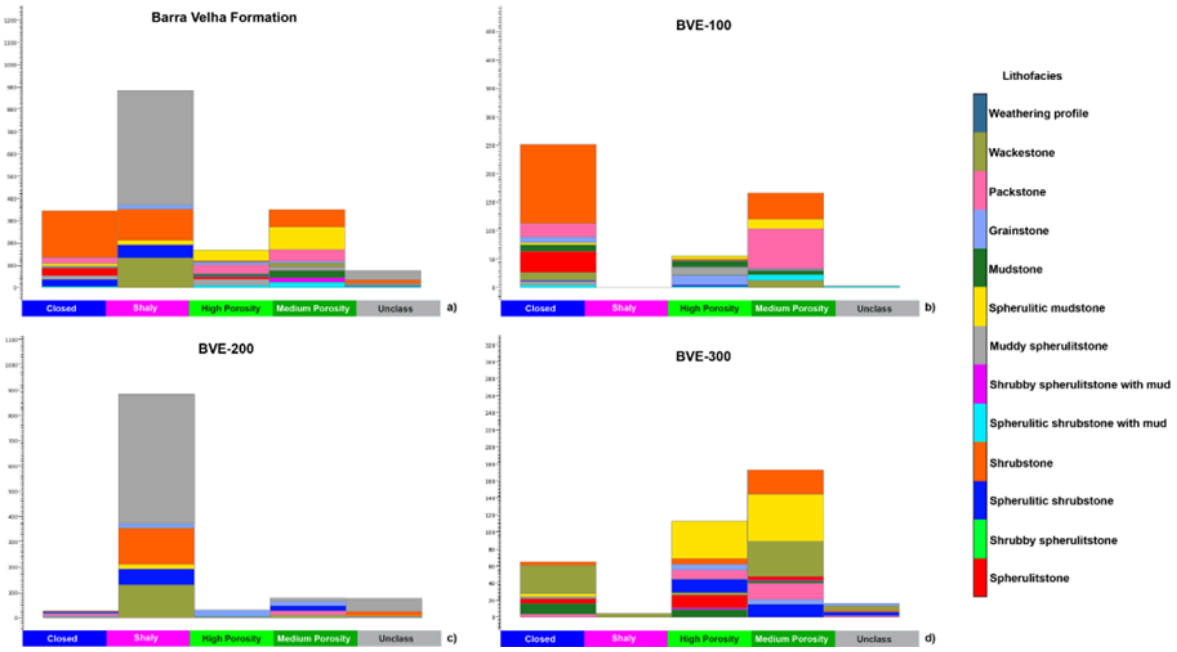

Figura 37: Histogramas exibindo a proporção de litofácies da Formação Barra Velha, conforme interpretação feita por Rebelo et al. (2022) a partir da classificação proposta por Gomes et al. (2020), para cada fácie petroelástica definida por Mattos et al. (2022). Fonte: Mattos et al. (2022).

Nos modelos gerados neste trabalho, as fácies e unidades que indicam os melhores reservatórios (High e Medium Porosity; UF4 e UF3; e GMU3) apresentam maior proporção no intervalo BVE100. No BVE200, predominam os resultados de classificação de fácies não-reservatório (Shaly, UF1 e GMU1). No BVE300, os modelos indicam ocorrência de fácies/unidades reservatório e não-reservatório (Shaly, High e Medium Porosity; GMU1 e GMU3).

Em geral, esses resultados estão de acordo com a interpretação geológica da área, uma vez que as fácies e unidades mais favoráveis são identificadas no BVE100 e 300, e as menos favoráveis no BVE200, como verificado por Mattos et al. (2022) (Figura 38).

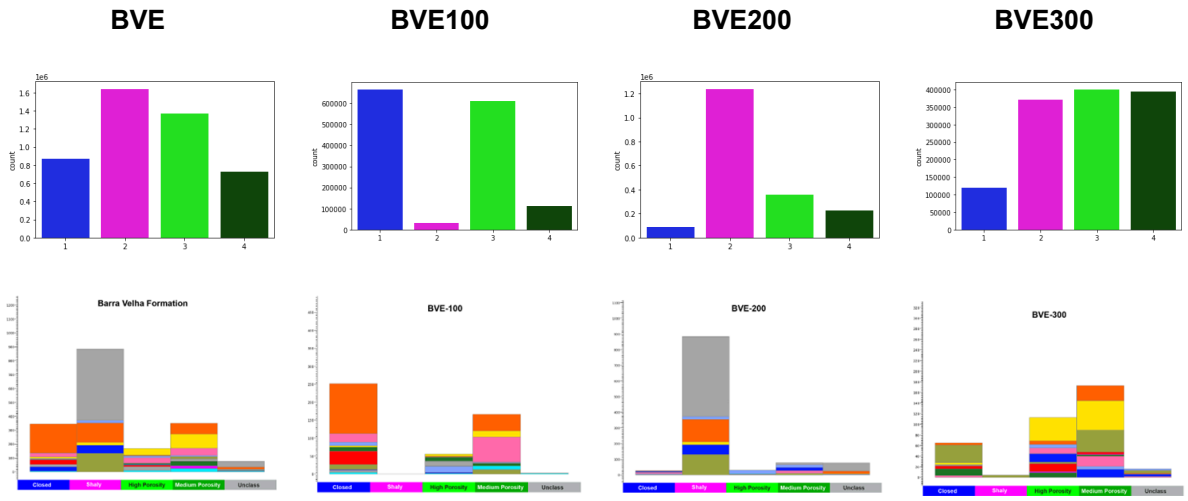

Figura 38: Histogramas apresentando a distribuição dos resultados de classificação de fácies petroelásticas dos modelos obtidos por este trabalho (primeira linha) e a proporção de litofácies da Formação Barra Velha, a partir da interpretação de Rebelo et al. (2022) com base na classificação de Gomes et al. (2020), para cada fácie petroelástica definida por Mattos et al. (2022). Histogramas da segunda linha extraídos de Mattos et al. (2022).

Com relação à distribuição espacial dos resultados de classificação da fácies petroelásticas, foi verificada a predominância de fácies argilosa nas porções mais baixas da área de estudo (linha pontilhada rosa da Figura 39). Além disso, foi identificado um padrão de ocorrência de fácies de alta porosidade na região mais elevada, especialmente no alto estrutural a NW (linha pontilhada verde da Figura 39), e predominância da fácies Tight na região mais elevada à SE (linha pontilhada azul da Figura 39).

A predominância de Tight na região mais elevada à SE pode ser explicada pelo processo de cimentação. Mattos et al. (2022) identificaram a predominância da litofácie *Shrubstone* na fácies Tight. Apesar desta litofácie consistir em fácies de granulação grossa, com estruturas em forma de leque que comumente preservam espaço poroso significativo entre elas, pode ocorrer cimentação de dolomita, afetando a porosidade dessas rochas (Mattos et al., 2022).

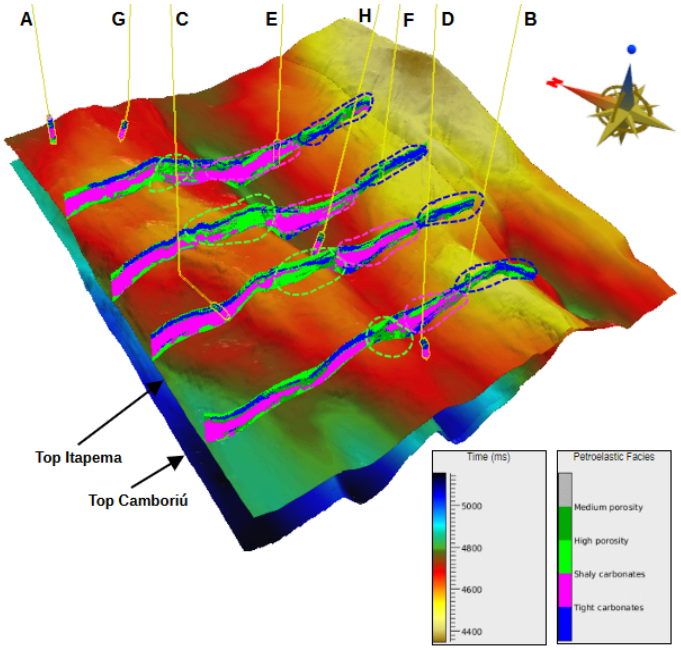

Figura 39: Crosslines apresentando o resultado de classificação das fácies petroelásticas.

Os resultados de classificação das unidades de fluxo também indicam uma predominância das unidades mais favoráveis para reservatórios na região mais elevada, especialmente no alto estrutural a NW (linha pontilhada amarela da Figura 40).

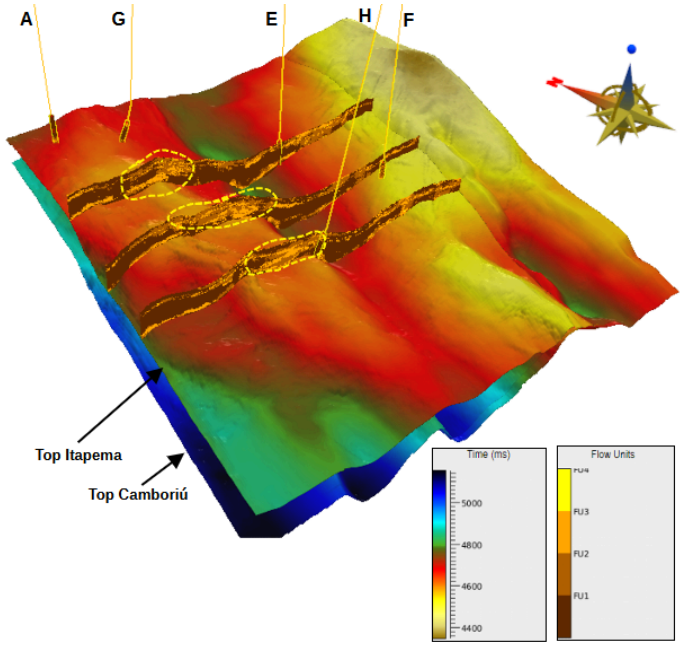

Figura 40: Crosslines apresentando o resultado de classificação das unidades de fluxo.

A partir da Figura 41, é possível observar um alinhamento da ocorrência da Unidade de Fluxo 4 com a direção do alto estrutural (NE-SW). Além disso, verifica-se uma correlação entre a fácies de alta porosidade (High Porosity) e a unidade de fluxo 4 (UF 4).

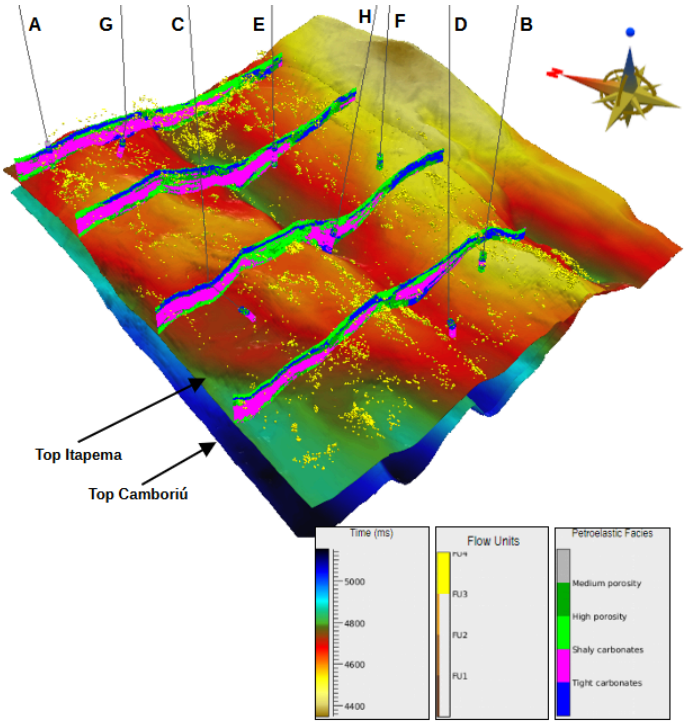

Figura 41: Crosslines apresentando o resultado de classificação das fácies petroelásticas e unidade de fluxo 4 (UF 4).

Santos (2022) interpretou 18 falhas de alto ângulo e com direção NE-SW na área do Tupi Nodes Pilot (eg. Cruz et al., 2021) (Figura 42). Nesse sentido, interpreta-se que os reservatórios ocorrem como corpos alongados na direção do alto estrutural, compartimentados em corredores entre falhas que ocorrem na mesma direção.

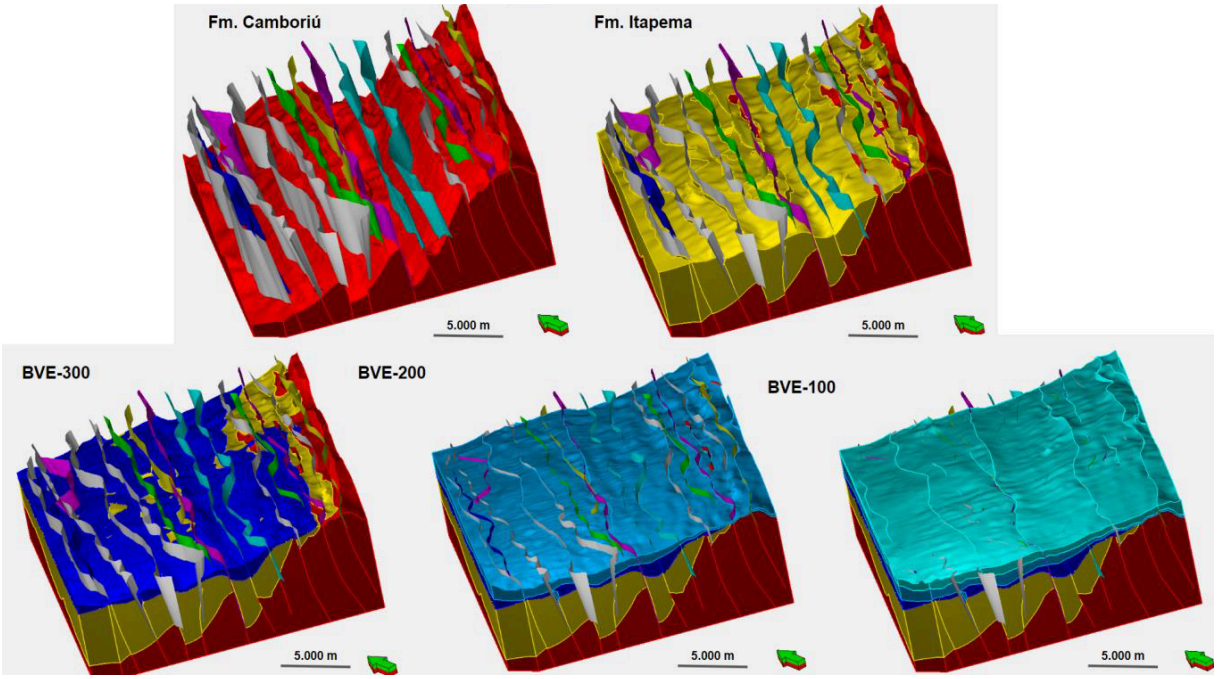

Figura 42: Modelo estrutural da área sudoeste do Campo de Tupi interpretado por Santos (2022) evidenciando as falhas e a compartimentação das zonas. Extraído de Santos (2022).

10. CONCLUSÕES

Esta pesquisa avaliou a aplicação do algoritmo XGBoost na classificação de dados categóricos. Nesse sentido, foi desenvolvido um fluxo de trabalho para classificação de fácies petroelásticas, unidades de fluxo e unidades geomecânicas definidas em escala de poço por Mattos et al. (2022), Rebelo et al. (2022) e Rojas (2023), respectivamente. As classificações utilizando o algoritmo XGBoost contaram com acurácia geral entre 70% e 95%, apontando o potencial do uso dessa ferramenta.

A distribuição de fácies e unidades dos modelos gerados por este trabalho indicam uma predominância de rochas reservatório nos intervalos de produção da formação Barra Velha BVE100 e BVE300. No BVE200 verificou-se a predominância de fácies e unidades não-reservatório (Shaly, UF1 e GMU3). Essa distribuição é consistente com a interpretação geológica da área de estudo, conforme verificado por Mattos et al. (2022). Adicionalmente, por meio dos resultados de classificação das fácies petroelásticas e unidades de fluxo, integrados com mapeamento estrutural em escala sísmica (e.g. Santos, 2022; Mattos et al., 2022) interpreta-se que os reservatórios ocorrem como estruturas alongadas na direção do alto estrutural, compartimentadas em corredores entre falhas que ocorrem na mesma direção.

Verificou-se que a aplicação dos algoritmos de reamostragem melhoraram significativamente o desempenho dos modelos de fácies petroelásticas. Com relação às unidades geomecânicas, os algoritmos de reamostragem não levaram a uma melhora significativa dos resultados. Entretanto, para as unidades geomecânicas, os resultados sem a aplicação de um método de reamostragem já apresentava bons resultados de precisão, acurácia sensibilidade e F1-Score.

Por outro lado, os resultados da classificação da unidade de fluxo UF 4 indicam uma limitação do modelo em classificar dados com baixa amostragem. A aplicação das técnicas de balanceamento de dados baseadas na criação da dados sintéticos (*Random Over Sampler* e SMOTE), na redução dos dados reais (*Random Under Sampler* e *Near Miss*) e em modelos híbridos (SMOTEENN e SMOTETomek) não resultaram em uma melhora significativa dos resultados de classificação. Esse resultado aponta uma grande limitação do modelo, uma vez que na exploração de hidrocarbonetos é comum que as classes que indicam os melhores reservatórios sejam menos amostradas que as classes não reservatórios.

Por fim, verificou-se através do cálculo de importância de variáveis que os vínculos geológicos e as pseudocurvas que foram geradas no fluxo de trabalho aplicado nesta pesquisa contribuíram para uma melhoria da classificação, chegando a representar o principal ganho do modelo na divisão de folhas das árvores. Essas melhorias são significativas e, portanto, o uso do XGBoost com as pseudocurvas adicionais é recomendado para classificar propriedades petrofísicas ou elásticas para caracterização de reservatórios ou estudos de simulação de fluxo nos reservatórios carbonáticos do pré-sal da Bacia de Santos.

REFERÊNCIAS

ANP. (2024). Boletim Mensal da Produção de Petróleo e Gás Natural. Agência Nacional do Petróleo, Gás Natural e Biocombustíveis. Mês de Fevereiro.

ANP. (2018). Shapefile de dados. Disponível em < http://www.anp.gov.br/exploracao-e-producao-de-oleo-e-gas/dados-tecnicos/shape-file-de-dados>. Último acesso em 4 de novembro de 2020.

Bosch, M., Mukerji, T., & Gonzalez, E. F. (2010). Seismic inversion for reservoir properties combining statistical rock physics and geostatistics: A review. Geophysics.

Breiman, L. (1996). Bagging predictors. Machine learning, 24, 123-140.

Breiman, L. (2001). Random forests. Machine learning, 45, 5-32.

Chang, H. K., Assine, M. L., Corrêa, F. S., Tinen, J. S., Vidal, A. C., & Koike, L. (2008). Sistemas petrolíferos e modelos de acumulação de hidrocarbonetos na Bacia de Santos. Revista Brasileira de Geociências.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research, 16, 321-357.

Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (pp. 785-794).

Cruz, N. M., Cruz, J. N., da Costa, M. M., Urasaki, E. A., Teixeira, L. M., & Grochau, M. H. (2021, September). 4D Seismic Applied to Pre-Salt Carbonate Reservoirs: Challenges and Results from Tupi Pilot, Santos Basin. In Second EAGE Conference on Pre-Salt Reservoir (Vol. 2021, No. 1, pp. 1-5). European Association of Geoscientists & Engineers.

Cypriano, L., Z. Yu, D. Ferreira, B. Huard, R. Pereira, F. Jouno, A. Khalil, et al, (2019). OBN for pre-salt imaging and reservoir monitoring — Potential and road ahead : Presented at the 16th International Congress of the Brazilian Geophysical Society, https://doi.org/10.22564/16cisbgf2019.318.

Detomo, R., Quadt, E., Pirmez, C., Mbah, R., & Olotu, S. (2012, November). Ocean bottom node seismic: learnings from Bonga, Deepwater Offshore Nigeria. In SEG International Exposition and Annual Meeting (pp. SEG-2012). SEG.

Doyen, P. (2007). Seismic reservoir characterization: An earth modelling perspective. Houten: EAGE publications.

Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. Journal of computer and system sciences, 55(1), 119-139.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. Annals of statistics, 1189-1232.

Gomes, J. P. Bunevich, R. B., Tedeschi, L. R., Tucker, M. E., Whitaker, F. F. (2020). Facies classification and patterns of lacustrine carbonate deposition of the Barra Velha Formation, Santos Basin, Brazilian Pre-salt. Marine and Petroleum Geology.

Hampson, D. (2010). Lithology prediction using seismic inversion attributes. CGG-Veritas, technical report.

Igual, L., & Seguí, S. (2017). Introduction to data science (pp. 1-4). Springer International Publishing.

Ivanhoe, L. F., & Leckie, G. (1993). Global oil, gas fields, sizes tallied, analyzed. Oil and Gas Journal.

Japkowicz, N. (2000). The class imbalance problem: Significance and strategies. In Proc. of the Int'l Conf. on artificial intelligence (Vol. 56, pp. 111-117).

Kearey, P., Brooks, M., & Hill, I. (2002). An introduction to geophysical exploration (Vol. 4). John Wiley & Sons.

Kubat, M. (2017). An introduction to machine learning. Cham, Switzerland: Springer International Publishing.

Leevy, JL. Khoshgoftaar, TM. Bauder, RA. Seliya, N. (2018). A survey on addressing high-class imbalance in Big Data. J Big Data; 5(1):42.

Ma, M., Zhao, G., He, B., Li, Q., Dong, H., Wang, S., & Wang, Z. (2021). XGBoost-based method for flash flood risk assessment. Journal of Hydrology, 598, 126382.

Mahesh, B. (2020). Machine learning algorithms-a review. International Journal of Science and Research (IJSR).[Internet], 9, 381-386.

Mani, I., & Zhang, I. (2003). kNN approach to unbalanced data distributions: a case study involving information extraction. In Proceedings of workshop on learning from imbalanced datasets (Vol. 126, No. 1, pp. 1-7). ICML.

Mann J. & Rigg J.W.D. (2012). New geological insights into the Santos Basin. GeoExpro.

Manzoor, U., Ehsan, M., Radwan, A. E., Hussain, M., Iftikhar, M. K., & Arshad, F. (2023). Seismic driven reservoir classification using advanced machine learning algorithms: A case study from the lower Ranikot/Khadro sandstone gas reservoir, Kirthar fold belt, lower Indus Basin, Pakistan. Geoenergy Science and Engineering, 211451.

Mattos, N. H., Rebelo, T. B., Leite, E. P., Cataldo, R. A., Batezelli, A. (2022). Quantitative Seismic Interpretation of the Barra Velha Formation in the Santos Basin, Se Brazil. Available at SSRN: https://ssrn.com/abstract=4237718 or http://dx.doi.org/10.2139/ssrn.4237718.

Mohriak, W. U. (2003). Bacias sedimentares da margem continental Brasileira. Geologia, tectônica e recursos minerais do Brasil.

Mohriak W, Nemčok M, Enciso G. (2008). South Atlantic divergent margin evolution: rift-border uplift and salt tectonics in the basins of SE, Brazil. Geological Society, London.

Moreira, J. L. P., Madeira, C. V., Gil, J. A., & Machado, M. A. P. (2007). Bacia de Santos. Boletim de Geociencias da Petrobras.

Nieto, J., Delbecq, F., and Batlai, B., (2011). Seismic Lithology Prediction – A Montney Shale Gas Case Study, CSEG CWLS Convention, Calgary, Alberta.

Normando, M. N., do Nascimento Junior, D. R., de Souza, A. C. B., Oliveira, K. M. L., Nepomuceno Filho, F., da Silva Barbosa, T. H., ... & de Almeida, N. M. (2022). A proposal for reservoir geostatistical modeling and uncertainty analysis of the Curimã Field, Mundaú Sub-Basin, Ceará Basin, Brazil. Journal of South American Earth Sciences, 114, 103716.

Penna, R., & Lupinacci, W. M. (2021). 3D modelling of flow units and petrophysical properties in Brazilian presalt carbonate. Marine and Petroleum Geology, 124, 104829.

Petrobrás. (2010). Declaração de comercialidade das áreas de Tupi e Iracema.

Pyrcz, M. J., & Deutsch, C. V. (2014). Geostatistical reservoir modeling. Oxford university press.

Rebelo, T. B., Batezelli, A., Mattos, N. H. S., & Leite, E. P. (2022). Flow units in complex carbonate reservoirs: A study case of the Brazilian pre-salt. Marine and Petroleum Geology, 140, 105639.

Riccomini, C., Sant, L. G., & Tassinari, C. C. G. (2012). Pré-sal: geologia e exploração. Revista Usp.

Santos, E. C. (2022). Mapeamento estrutural em escala sísmica e análise do deslocamento de falha no reservatório pré-sal do Campo de Tupi, Bacia de Santos. Trabalho de Conclusão de Curso (graduação) – Universidade Estadual de Campinas, Instituto de Geociências.

Schapire, R. E. (2003). The boosting approach to machine learning: An overview. Nonlinear estimation and classification, 149-171.

Skiena, S. S. (2017). The data science design manual. Springer.

XGBoost Developers. (2022). XGBoost Documentation. Disponível através do link: https://xgboost.readthedocs.io/en/stable/index.html. Último acesso em março de 2024.

Wang, X.; Wu, C.; Guo, Y.; Meng Q.; Zhang Y. and Tao Y. (2013). Accumulation Feature of Lula Oilfield and Its Exploratory Implication for Pre-salt Reservoirs in Santos Basin, Brazil.

Wang, Z. H. E., Wu, C., Zheng, K., Niu, X., & Wang, X. (2019). SMOTETomek-based resampling for personality recognition. Ieee Access, 7, 129678-129689.

Wei Q, Dunbrack RL Jr. The role of balanced training and testing data sets for binary classifiers in bioinformatics. PLoS One. 2013 Jul 9;8(7):e67863. doi: 10.1371/journal.pone.0067863. PMID: 23874456; PMCID: PMC3706434.

Witten, I. H., Frank, E., Hall, M. A., Pal, C. J., & DATA, M. (2005, June). Practical machine learning tools and techniques. In Data Mining (Vol. 2, No. 4).

Wright V.P. & Barnett A.J. (2015). An abiotic model for the development of textures in some South Atlantic early Cretaceous lacustrine carbonates.

Zhao, L., Geng, J., Cheng, J., Han, D. H., & Guo, T. (2014). Probabilistic lithofacies prediction from prestack seismic data in a heterogeneous carbonate reservoir. Geophysics.

Zhao, C., Jiang, Y., & Wang, L. (2022). Data-driven diagenetic facies classification and well-logging identification based on machine learning methods: A case study on Xujiahe tight sandstone in Sichuan Basin. Journal of Petroleum Science and Engineering, 217, 110798.