Pedro Vitor Abreu Affonso

MACHINE-LEARNING SEMI-SUPERVISIONADO APLICADO PARA PREDIÇÃO DE FÁCIES CARBONÁTICAS DA FORMAÇÃO BARRA VELHA NA BACIA DE SANTOS

Trabalho Final de Curso

(Geologia)

UFRJ
Rio de Janeiro
2022

Pedro Vitor Abreu Affonso

MACHINE-LEARNING SEMI-SUPERVISIONADO APLICADO PARA PREDIÇÃO DE FÁCIES CARBONÁTICAS DA FORMAÇÃO BARRA VELHA NA BACIA DE SANTOS

Trabalho de Conclusão de Curso de Graduação em Geologia do Instituto de Geociências, Universidade Federal do Rio de Janeiro – UFRJ, apresentado como requisito necessário para a obtenção do grau de Bacharel em Geologia.

Orientador(es):

Claudio Limeira Mello – UFRJ
Luciana Castro Brelaz – Petrec

Rio de Janeiro
NOVEMBRO/2022

Pedro Vitor Abreu Affonso

Machine-Learning Semi-supervisionado aplicado para predição de fácies carbonáticas da Formação Barra Velha na Bacia de Santos / Pedro Vitor Abreu Affonso - Rio de Janeiro: UFRJ / IGeo, 2022.

52p., 58 f

Trabalho Final de Curso (Geologia) – Universidade Federal do Rio de Janeiro, Instituto de Geociências, Departamento de Geologia, 2022.

Orientador(es): Claudio Limeira Mello, Luciana Castro Brelaz.

1. Geologia. 2. x – Trabalho de Conclusão de Curso. I. Claudio Limeira Mello. II. Universidade Federal do Rio de Janeiro, Instituto de Geociências, Departamento de Geologia. III. Machine-Learning Semi-supervisionado aplicado para predição de fácies carbonáticas da Formação Barra Velha na Bacia de Santos

Pedro Vitor Abreu Affonso

MACHINE-LEARNING SEMI-SUPERVISIONADO APLICADO PARA PREDIÇÃO DE FÁCIES CARBONÁTICAS DA FORMAÇÃO BARRA VELHA NA BACIA DE SANTOS

Trabalho de Conclusão de Curso de Graduação em Geologia do Instituto de Geociências, Universidade Federal do Rio de Janeiro – UFRJ, apresentado como requisito necessário para a obtenção do grau de Bacharel em Geologia.

Orientador(es):

Claudio Limeira Mello – UFRJ
Luciana Castro Brelaz – Petrec

Aprovada em: 14/12/2022

Por:

_____
Orientador: Dr. Claudio Limeira Mello, UFRJ

_____
Orientador: Dra. Luciana Castro Brelaz, Petrec

_____
Dr. Leonardo Borghi, UFRJ

_____
Dr. José Carlos Seoane, UFRJ

Agradecimentos

Vem chegando ao fim o primeiro capítulo da minha jornada geológica. Sou plenamente agradecido pelos pequenos meandros da vida que me transportaram até a sedimentação deste momento presente. Estes derradeiros anos de graduação puderam litificar na minha mente o significado do que é ser um geólogo. Através da busca em escutar as histórias que as rochas contam, aprendi que existem vastos mistérios encantadores escondidos no reino mineral que tanto tem a nos ensinar sobre humildade e reconhecimento. Hoje sei que, por debaixo de nossas botas enlameadas de cada pesquisa de campo que nos levaram a tantos cantos do Brasil durante a graduação, existem grandes forças construtivas esculpidas pela sensibilidade fina do tempo. Acredito com convicção que este é um conhecimento que deve chegar a toda a humanidade, pois traz um sentimento profundamente transformador e capaz de nos unir.

Eu preciso, portanto, agradecer a cada um dos amigos feitos durante este "caminho de pedras perdido na serra por onde não vai mais ninguém". Agradeço a todos os meus queridos amigos da vida que estiveram comigo desde o início até a conclusão deste ciclo, me dando suporte, boas conversas e conselhos. Quero agradecer especialmente aos meus companheiros de graduação que tanto me entusiasmaram com o amor e a paixão com que falam sobre geologia, fundamentais para a minha caminhada. São eles Bernardo Leão, Eduardo Sartori, Gil Pedro, Julia Mascarenhas, João Pedro Carneiro, Luan Dias, Lucas Locatelli, Paulo Vinícius, Raphaela de Negri, Thauan Vaisman e tantos outros que tive o prazer de conviver durante os tempos de pesquisas de campo e momentos de confraternização no Diretório Joel Valença. Se carrego em mim um caminho de pedra no peito, é especialmente graças a estas pessoas que sinto que terei a honra de dizer para novas gerações de geólogos que foram pessoas com quem estudei junto.

Eu também quero deixar um agradecimento especial a todos os meus orientadores e colegas de trabalho, que tanto me ensinaram e acrescentaram academicamente e profissionalmente. Sou grato ao professor Gerson, que me abriu as portas para a iniciação científica através da hidrogeologia, a professora Katia Mansur, com quem trabalhei durante anos através de trabalhos de extensão e divulgação geocientífica, e ao professor Claudio Limeira, que me permitiu a realização deste trabalho e tanto se empenha pela didática e compromisso com o aprendizado dos estudantes de graduação, além de tantos outros que me proporcionaram grandes oportunidades de conhecimento e inspiração durante o meu processo de formação. Eu também sou especialmente grato a minha orientadora Luciana Brelaz, colega de profissão e trabalho na Petrec, que me deu toda a devida atenção e suporte para a realização deste trabalho de conclusão de curso, e que me transmitiu o seu entusiasmo pela geologia das rochas carbonáticas, que vem sendo um tema novo de grande aprendizado para mim.

Resumo

ABREU, Pedro Vitor. **Machine-Learning Semi-supervisionado aplicado para predição de fácies carbonáticas da Formação Barra Velha na Bacia de Santos**. Rio de Janeiro, Ano. 2022, 59 f. Trabalho Final de Curso – Departamento de Geologia, Instituto de Geociências, Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2022.

Existe uma disponibilidade cada vez maior de dados geocientíficos de exploração disponíveis na indústria de óleo e gás. Com isso, ferramentas auxiliares baseadas em dados se tornaram importantes para otimizar o ganho de informação geocientífica a partir destes dados e permitir processos de tomada de decisão mais rápidos e confiáveis. No entanto, o desenvolvimento destas tecnologias depende da padronização da forma destes dados e de suas metodologias descritivas, que muitas vezes divergem entre os geocientistas e as diversas fontes destes dados, que recorrentemente também provêm de diferentes escalas. A complexidade de reservatórios não-convencionais, como os do Pré-sal brasileiro, elevam estas dificuldades já existentes. Neste sentido, este trabalho avalia os resultados de uma metodologia de *Machine-learning* semi-supervisionado que foi aplicada nos calcários aptianos da Formação Barra Velha no Pré-sal da Bacia de Santos. A metodologia segue uma abordagem de *PU-learning* com a utilização do algoritmo *Random-forest* baseada em dados públicos de testemunhos geológicos, amostras laterais e perfis geofísicos de poço no intervalo correspondente a estas rochas da Formação Barra Velha. Um agrupamento de fácies carbonáticas foi fornecido por uma equipe de geocientistas e então reagrupado com base em descrições quantitativas, qualitativas e critérios deposicionais relacionados a estas amostras com o objetivo de adequar estes agrupamentos à entrada no algoritmo de *Machine-learning*. Para lidar com o fato de as amostras pertencerem às diferentes escalas e fontes dos dados, as amostras descritas em escala de testemunho são selecionadas como “rotuladas” e as demais são “não-rotuladas”, estabelecendo um critério de confiabilidade das descrições das amostras e que se adequa à forma semi-supervisionada de aprendizado de máquina. Métricas de avaliação do modelo gerado foram calculadas e analisadas, em paralelo a uma comparação com os resultados de um modelo tradicional supervisionado. Os resultados demonstraram um ganho expressivo de precisão geral do modelo (> 10%) em relação à metodologia supervisionada, e sugestões críticas baseadas no resultado foram propostos para execução em futuros trabalhos de pesquisa neste segmento.

Palavras-chave: Bacia de Santos, Barra Velha, Carbonatos, *Machine-Learning*, Semi-supervisionado.

Abstract

ABREU, Pedro Vitor. **Machine-Learning Semi-supervisionado aplicado para predição de fácies carbonáticas da Formação Barra Velha na Bacia de Santos**. [*Semi-supervised Machine-learning applied for prediction of carbonate facies from Barra Velha formation of Santos Basin*.] Rio de Janeiro, Ano. 2022, 59 f. Trabalho Final de Curso –Departamento de Geologia, Instituto de Geociências, Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2022.

There is an increasing availability of geoscientific exploration data for the oil and gas industry. Supporting data-driven tools have become important for the optimization and geoscientific information gain from this kind of data and thus allowing a fastest and more trustable decision making. Nonetheless, the development of this kind of technology depends on the standardization of the data and its descriptive methodologies, that many times diverges between the geoscientists and its many data sources, that recurrently comes from different scales of samples. The complexity of non-conventional reservoir, like the ones from brazilian pre-salt, increases those pre-existing difficulties. In this sense, this work evaluates the results of a semi-supervised Machine-learning methodology that was applied to the aptian carbonates of Barra Velha formation, from the Santos Basin pre-salt. This methodology follows a PU-learning approach with the utilization of the Random-forest algorithm based on public data from geological cores, side samples and geophysical data from the corresponding depths of the Barra Velha carbonates. A team of geoscientists provided a carbonate facies grouping, and this work regrouped it based on quantitative and qualitative descriptions, and in depositional criteria related for those samples, aiming to better utilize this data for Machine-learning. To deal with the fact that the samples belong for different scales and data-sources, the classified samples from geological cores were select as “labeled”, and the rest of it was defined as “unlabeled”, establishing a criteria for description of the samples and that fits the workflow for semi-supervisioned Machine-learning. Model evaluation metric were analyzed and compared to results of a regular supervisioned model approach. The results show that the overall precision of the semi-supervisioned model has increased significantly by 10% in relation to the supersivioned methodology, and critical suggestions were made based on the results for motivation of future researches from this topic.

Key-Words: Santos Basin, Barra Velha, Carbonates, Machine-learning, Semi-supervisioned.

LISTA DE TABELAS:

LISTA DE FIGURAS:

Sumário

1. INTRODUÇÃO

Uma importante finalidade das atividades exploratórias da indústria de óleo e gás consiste em caracterizar reservatórios de hidrocarbonetos para otimizar o seu processo de produção. Para isso, a identificação das regiões com as melhores condições permo-porosas para o escoamento de fluidos depende da caracterização de fácies sedimentares para gerar modelos geológicos, fornecendo um forte tópico de pesquisa e desenvolvimento dentro do contexto que envolve a geologia do petróleo. Estas demandas são elevadas pelas condições do domínio do Pré-sal brasileiro: os desafios da exploração em reservatórios não-convencionais de águas ultra profundas exigem inovações tecnológicas e científicas que permitam o melhor aproveitamento da quantidade cada vez maior de dados disponíveis para análise.

Dado o grau de complexidade da matriz das rochas carbonáticas e as distribuições das propriedades petrofísicas associadas à deposição e ao histórico diagenético, a descrição de fácies sedimentares exige uma grande quantidade de tempo investido para a classificação através do conhecimento técnico de especialistas, tanto para as escalas de descrições das amostras obtidas em métodos diretos, como os testemunhos de sondagem, quanto aquelas descritas a partir de métodos indiretos, tais como os perfis geofísicos de poços. Os testemunhos de sondagem são limitados devido ao alto custo de aquisição, definindo uma baixa disponibilidade de amostras para métodos diretos de descrição faciológica. Em escala de perfis geofísicos de poço, existem mais dados disponíveis para interpretação, e uma variedade de combinações de perfis convencionais é efetivamente utilizada para identificar fácies carbonáticas, o que constitui um dos métodos mais efetivos para a caracterização de reservatórios (Momeni et al., 2019). No entanto, essas identificações dependem parcialmente do conhecimento técnico de profissionais especialistas e podem ser comprometidas devido à complexidade e à diferença dos dados dos perfis de poço em diferentes blocos de exploração, o que dificulta o estabelecimento de uma boa relação não-linear e universal entre os parâmetros de caracterização de reservatório e os resultados finais de processos de interpretação de fácies carbonáticas (X. Lan et al., 2021). Além disso, existem subjetividades intrínsecas ao interpretador que promovem a dificuldade de padronização de resultados.

Neste sentido, ferramentas que possam utilizar abordagens *data-driven* para otimizar o ganho de informação da grande quantidade de dados quantitativos dos métodos geofísicos e conciliar dados qualitativos das descrições geológicas surgem como uma possibilidade para obter informações auxiliares e facilitar tomadas de decisões em tempo real na indústria. Para estes tipos de problemas, algoritmos de *Machine-learning* são comumente utilizados para explorar informações contidas em conjuntos de dados e realizar tarefas de regressão ou classificação.

Métodos supervisionados de *Machine-learning* são os métodos mais amplamente utilizados para as tarefas de automatização de identificação de fácies sedimentares a partir de dados de poços (X. Lan *et* al., 2021). Esta metodologia consiste em extrair informações a partir de conjuntos de dados "rotulados" (classificados) para que um determinado algoritmo - como *Random Forest*, *Support Vector Machine*, entre outros - possa identificar padrões estatísticos e então identificar determinadas classes. Para o algoritmo realizar a tarefa de "aprender", os dados precisam estar previamente classificados, por isso é designado o termo "supervisionado". Além da necessidade de tempo necessário para a identificação de um grande número de amostras com suas respectivas fácies classificadas para estabelecer um conjunto de treinamento, existe um problema inerente ao balanceamento da frequência de ocorrência entre cada classe. A acurácia do modelo supervisionado depende de um conjunto de dados de treino com uma quantidade de classes balanceada, o que torna o poder de generalização dos modelos supervisionados bastante comprometido quando é aplicado a um contexto geocientífico de predição de fácies - que obedecem modelos deposicionais e, portanto, raramente ocorrem em mesma proporção ao longo de um perfil de poço.

Métodos "não-supervisionados", por sua vez, utilizam dados "não-rotulados" e buscam encontrar padrões estatísticos no conjunto de dados, sem atribuir o resultado a uma classe. Estes padrões estatísticos podem ser baseados em distâncias entre pontos, como o algoritmo *K-means*, em densidade de pontos, como o *Density Based Scan* (DBSCAN), entre outros. Segundo o trabalho publicado por Jain et al. (2019), dados petrofísicos não são bons para tarefas de "clusterização" (métodos não-supervisionados), pois os dados geocientíficos não são distribuídos esfericamente, e não são independentes entre uma amostra e outra, uma vez que são dependentes dos valores de profundidade, obedecendo critérios naturais de ordem de deposição.

Machine-learning semi-supervisionado é um estado intermediário entre o aprendizado supervisionado e o não-supervisionado (Jordan e Mitchell, 2015). A ideia de um aprendizado semi-supervisionado decorre da utilização de amostras rotuladas e não-rotuladas (classificadas e não-classificadas) para o reconhecimento de padrões (Chapelle et. al, 2006). A abordagem "*Positive and Unlabeled learning*" (*PU-learning*), proposta pelo trabalho de Elkan e Noko (2008), permite estabelecer um classificador distinto para cada classe (padrão "*one by one*") onde, para cada classificador, é calculada a probabilidade de uma amostra pertencer à uma classe.

Neste trabalho, uma metodologia semi-supervisionada de *Machine-learning* utilizando *PU-learning* permite integrar fácies carbonáticas descritas manualmente na escala de testemunho com os dados não-rotulados de perfis geofísicos convencionais no contexto petrofísico, assim permitindo um maior aproveitamento de todos os dados disponíveis, enquanto honra as descrições de especialistas - mesmo com uma menor quantidade de dados "rotulados".

São utilizados dados de 8 poços *offshore* no contexto dos carbonatos da Formação Barra Velha, do Pré-sal da Bacia de Santos, para agrupar fácies carbonáticas descritas em escala de testemunho em grupos que respeitem as respostas petrofísicas, estabelecendo critérios estatísticos que visem equilibrar as respostas quantitativas obtidas nos perfis geofísicos, com as descrições sedimentológicas e texturais identificadas nas rochas. Após o agrupamento de fácies, é feita uma comparação entre as métricas de resultados obtidos com a abordagem semi-supervisionada, e uma abordagem supervisionada

1.2 Objetivos

Este trabalho tem como principal objetivo desenvolver um modelo estatístico para classificação de agrupamentos de fácies carbonáticas da Formação Barra Velha, na Bacia de Santos. Para atingir esta finalidade, esta pesquisa deve conciliar desafios inerentes à problemas que surgem em projetos similares, tais como:

1. A necessidade de um conjunto de dados balanceado e padronizado.

2. A capacidade do modelo criado extrair o máximo de informações disponíveis que honrem tanto as distribuições petrofísicas (obtidas por medições geofísicas de perfis

convencionais) quanto as descrições texturais (realizadas em amostras de testemunhos de sondagem) de cada fácies.

3. A funcionalidade de realizar predições rápidas através de novos dados de entrada com boas métricas de precisão e acurácia gerais e individuais.

Neste sentido, uma metodologia de *Machine-learning* semi-supervisionado será aplicada através de uma abordagem de *PU-learning* com o algoritmo *Random-Forest*. Com isso, espera-se validar um fluxo de trabalho que permita suprir eficientemente as necessidades acadêmicas e industriais descritas acima, enquanto disponibiliza uma ferramenta capaz de otimizar processos exploratórios através da capacidade de gerar informações auxiliares para processos de interpretação e tomadas de decisão eficientes. Todo o fluxo de trabalho será desenvolvido através de códigos computacionais escritos em Python utilizando bibliotecas bem documentadas. O código final está disponibilizado publicamente no repositório do Github (https://github.com/pvabreu7/TCC) para que a metodologia possa ser transparente, reprodutível e aberta a críticas, seguindo princípios essenciais da metodologia e publicação científica.

2. CARACTERIZAÇÃO DA ÁREA DE ESTUDO

2.1 Localização

A área de estudo está situada na Bacia de Santos (Figura 1). A Bacia de Santos está localizada geograficamente ao longo do litoral dos estados do Rio de Janeiro, São Paulo, Paraná e Santa Catarina, limitando-se ao norte com a Bacia de Campos pelo Alto de Cabo Frio e, ao Sul, com a Bacia de Pelotas pela Plataforma de Florianópolis (Moreira et. al, 2007).

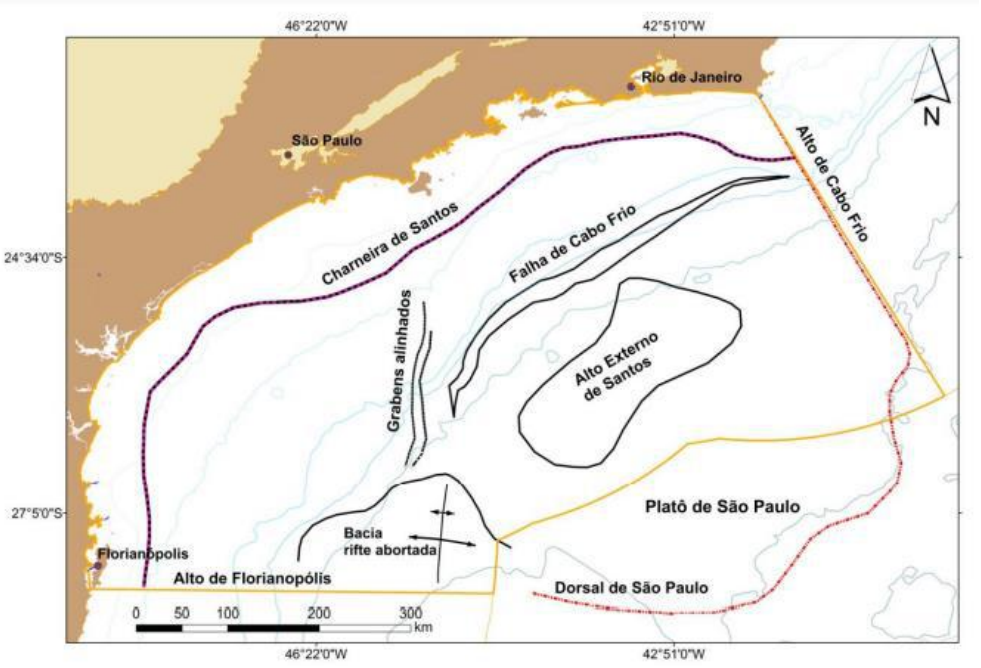

Figura 1: Mapa de localização da Bacia de Santos com os principais elementos do seu arcabouço regional (editado de Garcia *et* al., 2012)

A localização da área de estudo está inserida no contexto do Campo de Libra, cuja acumulação está localizada na porção Norte da Bacia de Santos e foi descoberta em 2010 pelo prospecto “Libra” (poço 2-ANP-2A-RJS), localizada em uma lâmina d’água de 1964m e profundidade final medida de 6023m (Petrobrás, 2016).

2.2 Geologia Regional

A Bacia de Santos é uma bacia de margem passiva de idade Neocomiana, sendo também a maior bacia sedimentar *offshore* do país, com uma espessura sedimentar de 10km em seu depocentro (Chang et. al, 2008) ao longo de uma área superior a 350.000 km² (Milani *et* al., 2000; Papaterra, 2010).

A origem e o desenvolvimento da Bacia de Santos estão diretamente relacionados à fragmentação do Supercontinente Gondwana durante o período Neocomiano, no Cretáceo Inferior, e posterior abertura e desenvolvimento do Oceano Atlântico. Este evento foi responsável pela separação dos continentes Sul-americano e Africano (Dias, 2004), que ocorreu em função dos esforços tectônicos de caráter distensivo ao longo de zonas de fraqueza pré-existentes do embasamento (Buckley *et* al., 2015).

O embasamento cristalino da bacia aflora em sua porção terrestre na região de São Paulo e é composto por granitos e gnaisses pré-cambrianos e metassedimentos da Faixa Ribeira, compondo o Complexo Costeiro (Moreira et al., 2007). O embasamento econômico da bacia é representado pelos basaltos da Formação Camboriú, de idade Neocomiana (andares locais Rio da Serra e Aratu), que ocorre discordante às rochas pré-cambrianas (Moreira et al., 2007). Ainda de acordo com Moreira et. al, 2007 o preenchimento da bacia pode ser dividido em três supersequências estratigráficas (Figura 2): sequências Rifte, Pós-Rifte e Drifte.

A fase da supersequência Rifte da bacia é composta pela Formação Camboriú, Formação Piçarras e Formação Itapema, que pertencem ao Grupo Guaratiba. Os basaltos da Formação Camboriú registram os grandes eventos magmáticos que desenvolveram extensos derrames toleiíticos que marcaram o início da fase Rifte da bacia e estão sotopostos a todo o preenchimento sedimentar da mesma. Esta formação é correlacionável com a Formação Serra Geral na Bacia do Paraná, segundo diversos autores. Durante a fase Rifte, ocorreu um processo de subsidência mecânica da bacia que persistiu até o Eoaptiano, quando as falhas mestras do rifteamento passaram a ter uma frequência cada vez menor de pulsos de atividade tectônica, até cessar quase completamente (Moreira et al, 2007).

O processo de subsidência mecânica da bacia condicionou, durante o Barremiano (andares locais Buracica sup. e Aratu), a deposição de leques aluviais de arenitos e conglomerados polimíticos constituídos de fragmentos de quartzo, feldspato e basalto nas porções proximais, e o desenvolvimento de paleoambientes de lagos profundos contendo arenitos de composição talco-estevensítica da Formação Piçarras, que ocorre discordante

ao topo da Formação Camboriú. Os sedimentos foram depositados de forma contínua desde o início da supersequência rifte até o estágio máximo de atividade de desenvolvimento dos hemi-grabens da bacia (Moreira et. al, 2007). Ao final da fase Rifte, a atividade tectônica relacionada à subsidência mecânica se torna mais rara, e passa a dar lugar a eventos de soerguimento regional do embasamento. Estas movimentações tectônicas foram responsáveis por criar altos estruturais intrabasinais que favoreceram o isolamento de áreas em relação ao aporte sedimentar de terrígenos de proveniência continental. É neste contexto que está inserida a Formação Itapema, de idade Neobarremiana à Eoaptiana (andares locais Buracica sup. ao Jiquiá) - depositada sobre uma discordância com o topo da Formação Piçarras (discordância intrabarremiana, 126Ma) – e caracterizada em termos de litoestratigrafia pela ocorrência de coquinas, grainstones e packstones bioclásticos, e folhelhos escuros ricos em matéria orgânica (Moreira et al., 2007). O topo desta formação coincide com a discordância regional pré-Alagoas (DPA), e encerra a supersequência Rifte.

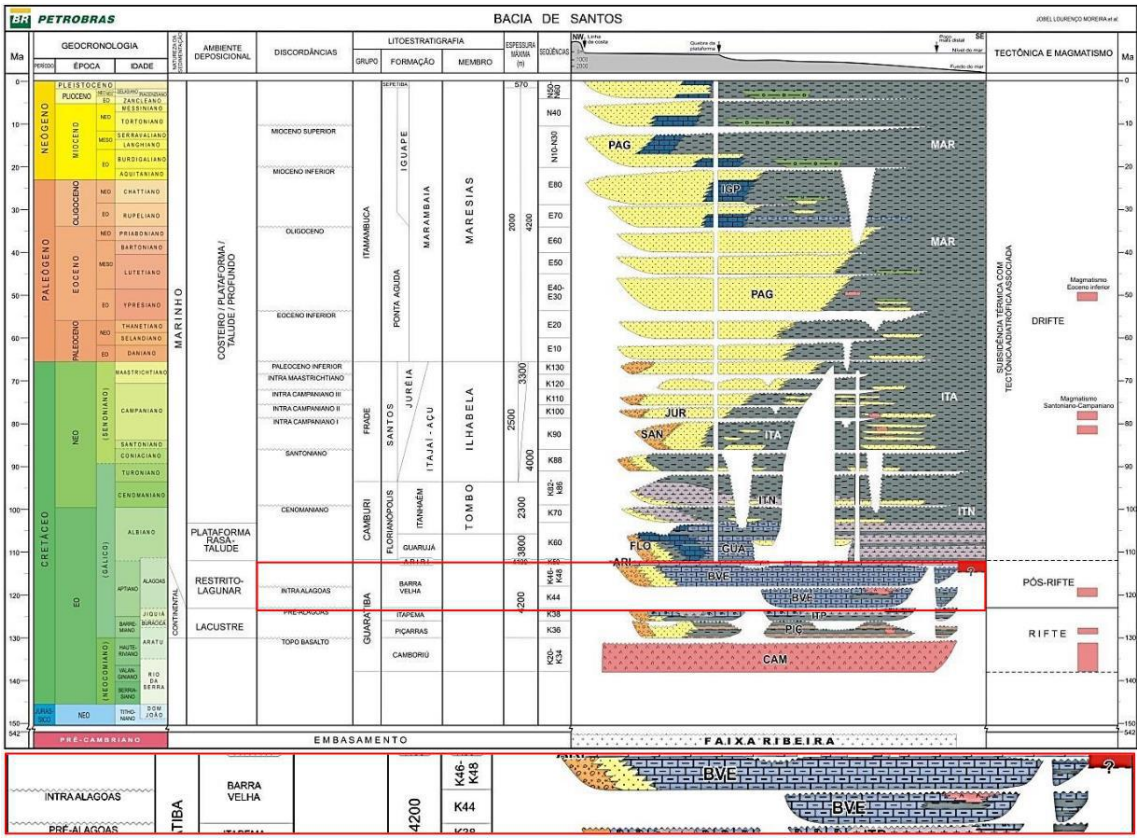

Figura 2: Carta estratigráfica da Bacia de Santos com destaque para o intervalo correspondente a Formação Barra Velha. Adaptado de Moreira et al., 2007.

Segundo Moreira et al. (2007), a supersequência pós-Rifte é representada pelos depósitos aptianos do grupo Guaratiba, que engloba as formações Barra Velha e Ariri.

Com a atenuação da atividade tectônica distensiva que desenvolveu os riftes, a evolução da bacia passou a ser dominada por subsidência termal, que por sua vez condicionou um estágio de bacia do tipo “sag”. Esta fase é caracterizada pelo desenvolvimento de extensas plataformas carbonáticas lacustres ao longo da margem sudeste brasileira e da margem oeste africana (Figura 3) durante o período Aptiano, correspondente ao andar local Alagoas (Gomes et al., 2009; Saller et al., 2016). Essas plataformas se formaram em condições de aridez, onde a alta evaporação associada a um paleoclima quente e seco, em conjunto com à contribuição de fontes hidrotermais, condicionou as características altamente salinas e alcalinas das águas dos lagos (Saller et al., 2016). As condições ambientais estressantes inibiram a ação de organismos predadores (ex: gastrópodes) e favoreceram a proliferação de organismos bioconstrutores como algas e cianobactérias, responsáveis pela formação dos grandes depósitos carbonáticos da Formação Barra Velha durante o Eoaptiano (Alagoas inferior). Importantes variações tectono-eustáticas ligadas aos processos geológicos de rifteamento e subsidência térmica da bacia sag estiveram diretamente ligadas as variações entre o crescimento de bioconstruções e a ocorrência de processos de carstificação e de exposição subaérea. Estas variações tiveram um impacto relevante na qualidade das fácies carbonáticas de reservatório (Dias, 2004; Gomes et al., 2012).

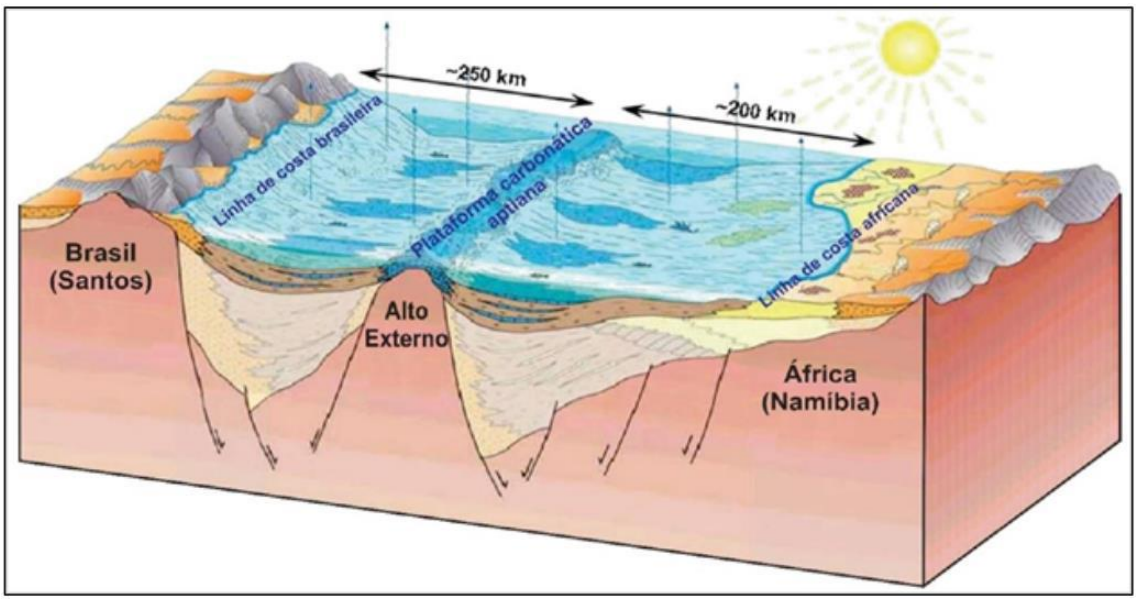

Figura 3: Modelo deposicional dos ambientes sedimentares nas margens conjugadas de Santos-Namíbia antes da deposição do sal e da ruptura do Gondwana. Fonte: Modificado de Gomes et al. (2012).

Litoestratigraficamente, a Formação Barra Velha foi subdividida por Moreira et al. (2007) em duas sequências deposicionais separadas por uma discordância de caráter regional em escala de bacia: inferior - de idade Eoaptiana, composta por

calcários/laminitos de origem microbial, estromatólitos e microbialitos dolomitizados total ou parcialmente e ricos em talco e argilas magnesianas nas fácies proximais e folhelhos carbonáticos nas porções distais, depositados sobre a discordância Pré-Alagoas – e superior, separada pela discordância Intra-Alagoas, de idade Neoaptiana, composta por leques aluviais conglomeráticos-areníticos nas porções proximais e de calcários estromatolíticos e laminitos microbiais, localmente dolomitizados nas porções distais da bacia (Ribeiro da Silva et al., 2021).

A presença de altos vulcânicos causou o confinamento das incursões marinhas, incrementando as condições de aridez que resultaram na formação de grandes bacias evaporíticas, que serviram como sítios deposicionais de espessas sequências de sal (predominantemente halita e anidrita) durante o Neoaptiano, que podem representar cerca de 2.000m de espessura, e são representadas pela Formação Ariri (Chang et al., 2008).

A derradeira supersequência é a da fase Drifte da bacia, caracterizada pelo espalhamento oceânico com desenvolvimento de crosta oceânica de margem passiva (Milani et al., 2000; Chang et al., 2008; Aslanian et al., 2009). A sedimentação é predominantemente marinha e relacionada à subsidência termal, que ocorreu a partir do Albiano até o recente. Pelo contexto de evolução do processo de subsidência termal da bacia, uma extensa plataforma carbonática de mar raso teve condições de ser implementada e se depositar sobre os evaporitos da Formação Ariri. Os carbonatos do Albiano foram posteriormente deformados pela movimentação dos evaporitos subjacentes a eles. A partir do Cenomaniano, os processos deposicionais passam a ser dominados por sedimentação siliciclástica, na medida em que a plataforma carbonática é progressivamente afogada. Sequências deposicionais de períodos de transgressão e regressão da linha de costa decorrentes das oscilações do nível relativo do mar caracterizam a predominância dos estratos da bacia até o recente.

2.3 Formação Barra Velha

A Formação Barra Velha foi individualizada como unidade geológica de idade Eoaptiano (andar Alagoas Inferior) a partir do poço 1-RJS-625 (Moreira et al., 2007), onde apresenta espessuras de mais de 300 m em contato concordante no topo com a base

da sucessão evaporítica da Formação Ariri (113 Ma) e discordante na base com o topo das coquinas da Formação Itapema através da discordância regional do pré-Alagoas (123,1 Ma). A Formação Barra Velha é cronocorrelata à Formação Macabu da Bacia de Campos e a sucessões carbonáticas aptianas da margem africana conforme descrito por Saller et al (2016) na Bacia de Kwanza.

Esta formação ocorre somente em subsuperfície ao longo de toda extensão após a charneira cretácica da bacia, sendo constituída por fácies predominantemente carbonáticas depositadas em relativa quiescência tectônica, onde as falhas mestras ativas durante a fase Rifte haviam cessado sua atividade, dando lugar a processos de subsidência termal. Estes depósitos englobam litofácies formadas *in situ* e litofácies retrabalhadas a partir dos depósitos plataformais formados in situ.

De acordo com Moreira et al. (2007), as principais litofácies formadas in situ são calcários estromatolíticos-esferulíticos, laminitos microbiais, microbialitos ricos em talco e argilas magnesianas (com esferulitos) e folhelhos carbonáticos presentes nas partes mais profundas e distais da bacia, depositados em ambiente deposicional transicional entre continental e marinho raso. Os depósitos de retrabalhamento incluem grainstones e packstones com fragmentos de estromatólitos, laminitos microbiais e esferulititos. Coquinas e basaltos datados em 118 Ma ocorrem de modo esporádico, intercalado aos microbialitos.

Em trabalhos mais recentes, a origem microbial dos carbonatos da Formação Barra Velha vem sendo refutada (Wright e Barnett, 2014). Assim, pesquisas foram publicadas sugerindo um modelo abiótico para o desenvolvimento das texturas carbonáticas dos carbonatos lacustres do cretáceo. O trabalho de Wright e Barnett (2014) identificou ciclotemas assimétricos que descrevem três fácies depositadas em função de um modelo geoquímico baseado em precipitação química controlada pelas condições de pH e atividade iônica das águas de um paleoambiente lacustre em um sistema de ciclos transgressivos-regressivos, sendo elas (Figura 4): carbonatos finos laminados acumulados na fase de afogamento, esferulitos de calcita em uma matriz de silicatos magnesianos e *shrubs* de calcita. Neste sistema, os eventos pluviométricos estão diretamente associados com a deposição de silicatos de Mg. A ação da pluviosidade causa a deposição de carbonatos laminados, expandindo os ambientes lacustres rasos da fase Rift da bacia. A evaporação ocorre em seguida, causando a precipitação dos silicatos de Mg e a nucleação de calcita em “géis” que produzem textura esferulítica. Quando a taxa de precipitação

destes “géis” diminui ou cessa completamente, a calcita pode precipitar com menor inibição, e formar *shrubs* análogos aos formados abióticamente em sistemas de travertinos modernos.

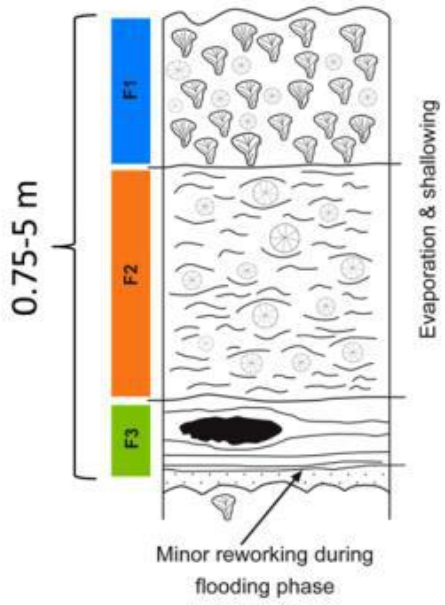

Figura 4: exemplo esquemático do ciclotema deposicional dos calcários da Formação. Barra Velha (Wright e Barnett, 2014). Fácies 1: *Shrubs* formados a partir de fluidos com baixa concentração de Mg/Ca; pouca influência microbial; boa porosidade intergranular. Fácies 2: esferulitos e estivensita; desenvolvimento de porosidade com a dissolução de estivensita. Fácies 3: fase de inundação; a redução da alcalinidade/salinidade provoca a precipitação de sílica com o decaimento do pH.

Neste sentido, os carbonatos descritos por Moreira et al. (2007) foram então redescritos por Arienti et al. (2018) e Souza et al. (2018) como: a) calcários cristalinos arborescentes-arbustiformes (shrubby boundstones); b) travertinos e tufas acamadados; e c) boundstones microbiais laminados (que incluem os laminitos microbiais). Os calcários cristalinos arbustiformes são compostos de agregados de cristais de calcita fascicular-ótica, associados a intercalações de argilitos silicáticos magnesianos como kerolita, estivensita e sepiolita. Os esferulitos estão frequentemente associados às argilas estivensíticas em fácies de argilitos laminados e wackestone-packestone esferulíticos. As rochas biogênicas são representadas por boundstones microbiais laminados (Souza et al., 2018).

Depósitos carbonáticos retrabalhados em condições de alta a baixa energia foram definidos como rudstones, grainstones (com grãos carbonáticos de intraclastos e esferulitos) e mudstones siltosos (Arienti et al., 2018). Fácies relacionadas a pedogênese e exposição subaérea são menos frequentes e incluem calcretes, dolocritos, silcretes e

brechas hidráulicas, com intensa substituição e/ou cimentação por dolomita e quartzo, além de dissolução.

A sedimentação dos depósitos da Formação Barra Velha ocorreu em estilos estruturais distintos. Os principais sistemas deposicionais incluem buildups carbonáticos (com calcário arborescentes-arbustiformes e travertinos), sistemas lacustres costeiros (shoreface, foreshore, berma/crista, leque de washover produzido por ação de tempestades), sistema lacustre distal, sistema de planície microbial, leques aluviais e sistema fandelta, sistemas evaporíticos argilosos magnesianos e sistemas palustres (Arienti et al., 2018).

3. MATERIAIS E MÉTODOS

Para a realização e a finalidade de atingir os objetivos deste trabalho, diversos tipos de dados foram analisados e reorganizados, quantitativamente e qualitativamente, a fim de alcançar um equilíbrio que honre os aspectos geológicos e petrofísicos do conjunto de dados. Estes materiais referem-se as amostras de rochas em escala de testemunho, amostras laterais e conjuntos de perfis petrofísicos convencionais, que estão apresentados na Tabela 1. A fonte dos dados é pública, e provém da Agência Nacional do Petróleo, Gás Natural e Biocombustíveis (*ANP), adquirida através da solicitação da Petroleum Research and Technology* (Petrec), empresa a qual um projeto de pesquisa e desenvolvimento para predição de fácies através de técnicas de *Deep Learning* deu origem e motivação a este trabalho de conclusão de curso.

Grupos de Perfis Petrofísicos	Mnemônicos
Gamma Ray + Gamma Ray Espectral	GR, POTA, URA, THO
Sônico	DTCO, DTSM, E, PR, VP, VS, VPVS
Triplo Combo	NPHI, RHOB, PEF
Resistividade	RESDEEP, RESMED, RESSHAL
Litogeoquímica	DWCA, DWAL, DWSI, DWFE, DWSU

Tabela 1: Conjuntos de perfis petrofísicos de poço disponíveis para este trabalho. GR: Gamma Ray. POTA, URA, THO: Potássio, Urânio e Tório. DTCO, DTSM: Delta-T Compressional e Delta-T Cisalhante. E: Módulo de elasticidade. PR: Razão de Poisson. VP, VS, VPVS: Velocidade das ondas compressionais (P), de cisalhamento (S) e razão entre VP e VS. NPHI, RHOB, PEF: Porosidade neutra, Densidade real, Fator fotoelétrico. RESDEEP, RESMED, RESSHAL: Resistividade profunda, média e superficial. DWCA, DWAL, DWSI, DWFE, DWSU: Densidades de Ca, Al, Si, Fe e S.

Os dados que compreendem as rochas dos testemunhos de sondagem e amostras laterais referem-se às amostras obtidas em 8 poços *offshore* da Bacia de Santos, em profundidades que testemunharam a Formação Barra Velha. São 470m de rochas testemunhadas em que 11 litofácies foram identificadas e descritas pela equipe de geocientistas da Petrec. As fácies das amostras laterais foram fornecidas pelos relatórios de descrição da ANP, e então incorporadas ao conjunto de dados original, mas não serão utilizadas no conjunto de dados reagrupado utilizado para este trabalho. A motivação

desta decisão está baseada na ideia de estabelecer um critério de confiança nas descrições e diminuir o erro associado à integração dos dados das amostras laterais com os perfis petrofísicos, que será discutido e explicado neste capítulo.

As fácies foram agrupadas em função de características texturais e petrofísicas em comum, de modo a simplificar o trabalho de treinamento de algoritmos de classificação supervisionada. Neste presente trabalho, é proposto um novo reagrupamento de fácies a partir do agrupamento estabelecido por esta equipe de geocientistas. O reagrupamento foi feito a partir das descrições das fácies das rochas de testemunhos em conjunto com análises estatísticas em boxplots e gráficos de barra. As fácies identificadas nos testemunhos, o agrupamento estabelecido pela equipe de geocientistas da Petrec, e o novo reagrupamento estabelecido neste trabalho são tópicos de apresentação e discussão deste trabalho.

Além das metodologias utilizadas para analisar e reagrupar as fácies carbonáticas, técnicas de pré-processamento também foram aplicadas para estabelecer um controle de qualidade dos dados. Primeiro, as fácies descritas em testemunhos de sondagem foram extrapoladas para a escala de perfis petrofísicos para integração entre as fontes de dados. Em seguida, fácies correspondentes às rochas ígneas e metamórficas foram removidas utilizando a metodologia de De Oliveira et al. (2019), que propõe uma abordagem para estabelecer parâmetros de "igneabilidade" para identificar intervalos que correspondem a rochas ígneas e metamórficas a partir de perfis de densidade, litogeoquímica e fator fotoelétrico. Por último, é aplicada uma separação entre os dados rotulados e não-rotulados para a abordagem de *Machine-learning* que será executada, onde os dados provenientes de amostras laterais e de perfis geofísicos são os dados "não-rotulados" (que não possuem classificação de fácies no conjunto de treino). O embasamento técnico e a argumentação para cada uma destas etapas de pré-processamento serão explicadas com maiores detalhes em seus respectivos índices.

Com os dados reagrupados o fluxo de trabalho de *Machine-learning* é executado: é definido o algoritmo de classificação, os conjuntos de treinamento e teste, os critérios de normalização dos dados e de validação das métricas de precisão e acurácia globais e individuais (para cada agrupamento de fácies). A abordagem semi-supervisionada utilizada é a de *PU-Learning*, com treinamento a partir do algoritmo *Random-forest*. Para cada agrupamento de fácies, um modelo de treinamento individual é definido. Através de cada um dos modelos treinados, é calculada a probabilidade de cada amostra pertencer a

um agrupamento de fácies e, então, é selecionado o agrupamento de fácies com a maior probabilidade. Para validação da metodologia, além dos cálculos de precisão e acurácia, também serão plotados gráficos de matriz de confusão e da distribuição de fácies predita em cada poço utilizando o conjunto de dados de teste, para a finalidade de comparar os resultados.

3.1 Pré-Processamento

A etapa de pré-processamento dos dados envolve a integração entre as diferentes escalas de amostras de rocha e amostras de perfis petrofísicos (integração rocha-perfil) e o controle de qualidade dos dados para que sejam filtrados os intervalos de profundidade que possuam apenas zonas de carbonatos limpos, assegurando a assertividade do modelo de predição de agrupamentos de fácies carbonáticas.

3.1.1 Amostras Laterais

Uma amostra lateral é um tipo de amostra comumente utilizada em rotinas de trabalhos de exploração e produção de óleo e gás e que possui um excelente custo-benefício. Através de ferramentas de perfilagem, é possível extrair partes de uma formação geológica que estão localizadas nas áreas adjacentes de um poço já perfurado e fornecer amostras em profundidades específicas que podem conter informações cruciais para estudos geológicos e de engenharia destas zonas.

Dados de amostras laterais foram fornecidos pela ANP e incorporados no conjunto de dados utilizados pela Petrec para descrição de fácies geológicas e posterior utilização em uma abordagem de classificação de Machine-learning supervisionada. Existem, no entanto, dois problemas principais na integração deste tipo de dado com os demais (dados de testemunhos e os perfis petrofísicos): o primeiro é a diminuição da acurácia das descrições geológicas – que por possuírem um fator de viés pessoal na interpretação, podem não ter seguido os mesmos padrões das descrições da equipe de geocientistas da Petrec em comparação com as descrições manuais feitas nos testemunhos, e o segundo é o erro inerente à calibração das profundidades das amostras laterais com os perfis petrofísicos. Apesar dos problemas apresentados, a utilização destes

dados faz sentido para uma abordagem de Machine-learning supervisionado, uma vez que adicionam mais dados classificados no conjunto de treino, diminuem o desbalanceamento da frequência de ocorrência de cada fácies, e assim permitem um melhor aprendizado do algoritmo.

Uma das grandes vantagens da abordagem de Machine-learning semi-supervisionado é o aproveitamento de todas as amostras sem o comprometimento da acurácia causado pelos problemas citados em função da integração de diferentes fontes de dados. Este tipo de abordagem não necessita de um conjunto de dados inteiramente classificado e, portanto, é possível permanecer com os dados das profundidades em que ocorrem as amostras laterais, sem associar uma classificação a elas.

Por estes motivos, para este trabalho os dados de amostras laterais serão utilizados de maneira "não-rotulada", isto é, não classificada. Desta maneira, estas amostras permanecem contendo atributos petrofísicos relacionados a elas, mas sem possuir classificação de fácies. Assim é possível padronizar as descrições de fácies obtidas pelas amostras de testemunho em conjunto com os perfis petrofísicos, estabelecendo estas classificações como aquelas que possuem um bom grau de confiabilidade para o treinamento do algoritmo de Machine-learning.

3.1.2 Extrapolação dos Dados para Integração Rocha-Perfil

As fácies definidas na escala dos testemunhos foram ajustadas à escala de amostragem média das ferramentas de perfis de poço para a extrapolação rocha-perfil. Sabe-se que, em média, a maior resolução alcançada por uma ferramenta de perfil convencional é de 15 cm, onde abaixo disso não é possível distinguir ou captar variações nas leituras.

Portanto, o valor de 15 cm foi convencionado como espessura mínima para delimitação de uma fácies no testemunho para correlação rocha-perfil. Neste sentido, todas as camadas sedimentares menores que 15 cm foram inseridas em outras fácies adjacentes com espessuras superiores a 15 cm, tendo em vista que laminações ou camadas muito finas estão em sub-resolução para perfis de poços. Além do redimensionamento das fácies, foi necessário correlacionar fácies e suas respectivas respostas físicas nos

perfis, para checar a existência de padrões de propriedades físicas e petrofísicas de fácies na escala de poço.

3.1.3 Identificação de Intervalos de Carbonatos Limpos

Para a identificação dos diferentes litotipos não-carbonáticos que ocorrem intercalados aos reservatórios carbonáticos do Pré-sal, como rochas siliciclásticas e ígneas, foi utilizada uma metodologia proposta por De Oliveira et al. (2019). Essa metodologia consiste na observação da variação do comportamento dos perfis geofísicos de densidade, fator fotoelétrico, e litogeoquímicos (Ca, Al, Si e Fe) em relação a cada litologia.

Parâmetros litológicos e composicionais como a Igneabilidade, Fe-Al-Ca e Si-Ca foram criados para serem plotados como curvas em escalas apropriadas, e então foi possível perceber a separação visual das litologias. Deste modo, foram observadas respostas específicas para as seguintes litologias: rochas ígneas máficas, rochas carbonáticas com e sem argila, rochas siliciclásticas (folhelhos, siltitos), rochas carbonáticas com metamorfismo de contato (rochas carbonáticas muito silicificadas) e anidrita. A geração destes perfis litológicos permitiu individualizar de modo mais assertivo o intervalo de interesse (carbonatos limpos) dos reservatórios, excluindo as profundidades onde ocorrem as litologias não-carbonáticas.

3.2 Descrição e Agrupamento das Fácies Carbonáticas

O bom agrupamento de fácies carbonáticas é essencial para que o algoritmo de Machine-learning tenha sucesso na etapa de treinamento. Para que a máquina possa entender e diferenciar fácies carbonáticas, é necessário que estes agrupamentos sejam suficientemente distintos entre si em função de critérios estatísticos. Para este trabalho, os critérios estatísticos representam distribuições de elementos petrofísicos que caracterizam cada fácies individualmente.

Neste sentido, um conjunto de dados originais é fornecido pela equipe de geocientistas da Petrec. Estes dados originais possuem 11 fácies carbonáticas

identificadas em escala de poço que compreendem 8 agrupamentos de fácies que levam em conta respostas registradas pelas amostras em perfis petrofísicos e as características texturais e mineralógicas descritas nas rochas a partir dos testemunhos e os relatórios de amostras laterais. Estes agrupamentos de fácies individuais podem ser observados na Tabela 2. As descrições de cada fácies individual e os seus respectivos agrupamentos serão apresentados a seguir.

Sigla do agrupamento	Fácies em escala de poço
MUD	Mudstone
MUD-arg	Mudstone argiloso
ESF-ARB	Calcário esferulítico-arbustiforme
ARB-ESF	Calcário arbustiforme-esferulítico
ARB-ARS	Calcário arbustiforme-arborescente
ARB-ARS-cao	Calcário arbustiforme-arborescente caótico
GST-RUD	Grainstones, rudstones
GST-RUD-cao	Grainstones e rudstones caóticos

Tabela 2: Agrupamento de fácies original, fornecido por descrições da Petrec.

3.2.1 Conjunto de Dados Originais

O conjunto de dados originais possui um total de 2296 amostras, divididas em 8 classes que correspondem a agrupamentos de fácies carbonáticas. A frequência de ocorrência entre elas ocorre de maneira desbalanceada (Figura 5).

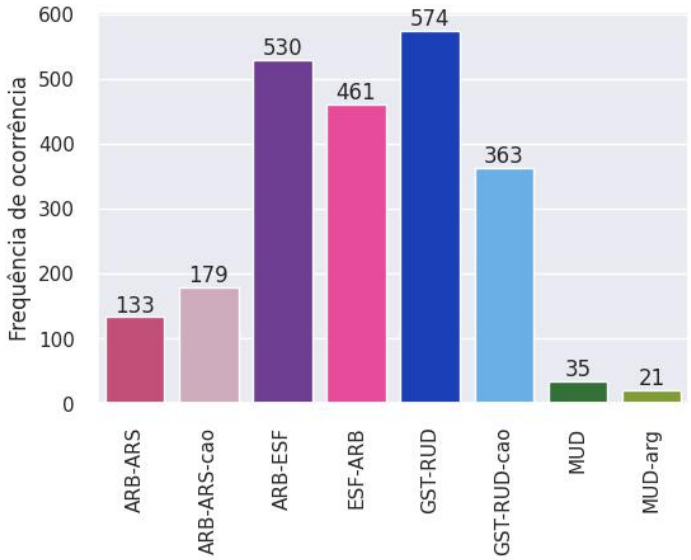

Figura 5: Frequência de ocorrência do agrupamento de fácies original.

3.2.1.1 Mudstone (MUD) e Mudstone argiloso (MUD-arg)

Mudstone é uma classificação de rocha carbonática cujo termo foi proposto por Dunham (1962). Este termo é designado para classificar uma rocha carbonática suportada por matriz com menos de 10% de grãos de tamanho areia ou menor.

Existem dois agrupamentos de fácies que caracterizam ocorrências de mudstones testemunhados no intervalo da Formação Barra Velha nos poços utilizados para este trabalho: MUD e MUD-arg. A motivação de diferenciar os mudstones entre estes dois agrupamentos baseia-se nas diferentes respostas obtidas em perfis petrofísicos que medem o teor de argilosidade destas rochas, como os perfis de raios gamma (SGR) e raios gamma espectral (THO, U e POTA). A distribuição dos agrupamentos de fácies em função destes perfis pode ser observada na Figura 6, onde é possível destacar o caráter argiloso da fácies MUD-arg, cuja mediana possui valores expressivamente maiores com relação às demais classes.

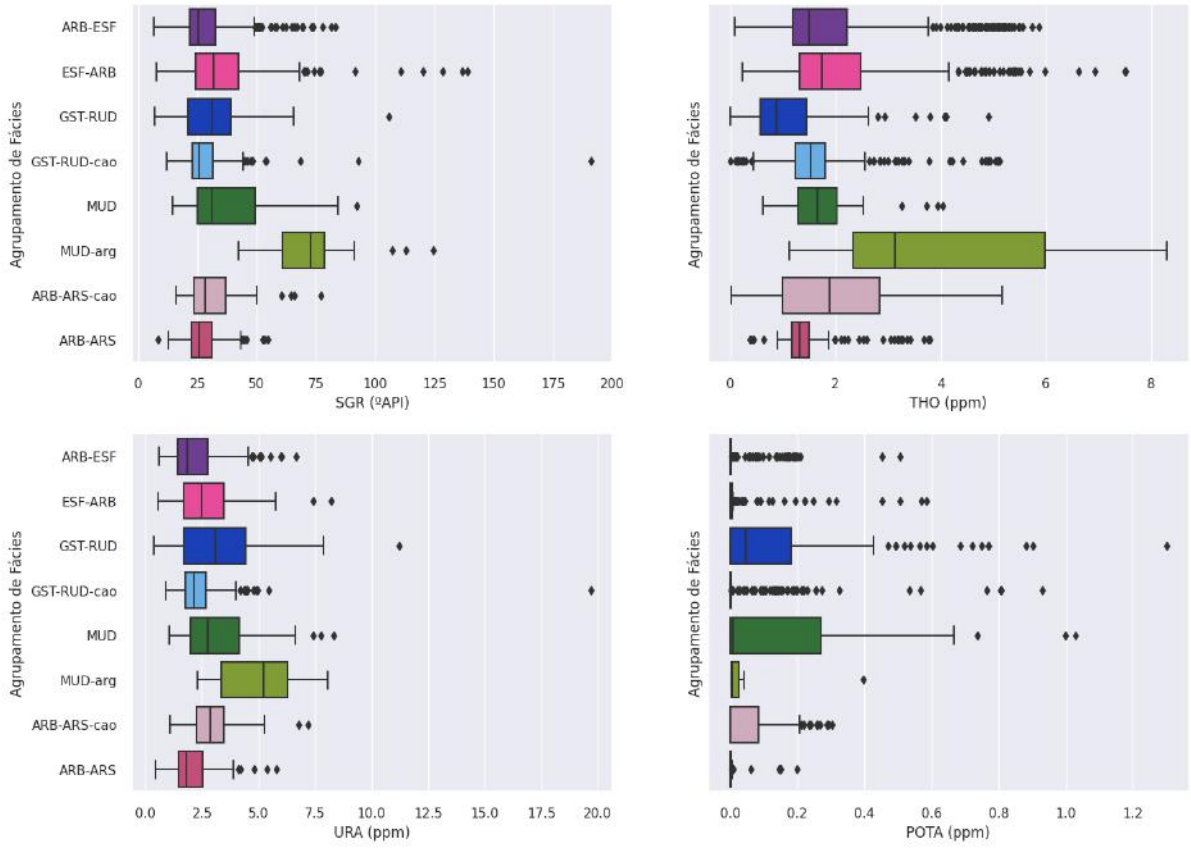

Figura 6: Distribuições dos perfis de raios gamma e raios gamma espectral para cada agrupamento de fácies.

Nos perfis petrofísicos, a fácies MUD é caracterizada principalmente por possuir valores baixos de argilosidade (perfis GR, THO, U, K), baixa a moderada porosidade (<10%) e altíssima resistividade (>300 ohm). Nas amostragens dos testemunhos geológicos, nota-se que ocorrem mudstones intercalados com calcários arbustiformes e esferulititos, litologias que correspondem aos agrupamentos de fácies ARB-ESF e ESF-ARB.

O agrupamento de fácies MUD-arg, por sua vez, é caracterizado por possuir respostas moderadas a altas em perfis sensíveis a argilosidade (como GR, THO, U e K) e baixa resistividade (<100 ohm). Nas amostras de rocha, ocorrem associados a brechas calcárias de matriz argilosa.

3.2.1.2 Calcário Esferulítico-Arbustiforme (ESF-ARB) e Arbustiforme-Esferulítico (ARB-ESF)

Os agrupamentos ESF-ARB e ARB-ESF compreendem as litologias calcário arbustiforme e esferulitito. Ambas litologias podem ocorrer de forma maciça ou laminada, e estão também presentes nestes agrupamentos. A principal motivação pela divisão entre estes agrupamentos é a predominância de cada fácies, em que ARB-ESF ocorre com uma frequência maior de calcários arbustiformes e, na ESF-ARB, com maior predominância de esferulititos.

De acordo com Terra et al. (2010), um estromatólito arbustiforme corresponde à uma rocha carbonática de origem frequentemente microbial cuja estrutura é laminada e, muitas vezes, convexa, que pode apresentar estruturas de crescimento da base para o topo. Quando as estruturas internas se organizam de forma ramificada ou não desde a base e a razão entre a altura e a largura é de aproximadamente 1:1, é designado o termo “arbustiforme”. De acordo com publicações científicas mais recentes, estas rochas carbonáticas que ocorrem na Formação Barra Velha são chamadas de calcários arbustiformes (Arienti et al., 2018; Souza et al., 2018) ao invés de “estromatólitos”, que sugerem uma origem microbial. Estes nomes tornam-se mais apropriados ao estar em conformidade com trabalhos recentes que apontam as similaridades entre o desenvolvimento das texturas destes carbonatos com os travertinos formados a partir da

ação de fontes hidrotermais, sugerindo uma origem abiótica para estas rochas (Wright e Barnett, 2014; Muniz e Bosence, 2015; Wright e Barnett, 2017).

Nos testemunhos descritos, os calcários arbustiformes ocorrem de forma maciça ou laminada, com *shrubs* muito pequenos (< 2 mm) a médios (5 a 1,5 cm) e empacotamento denso a normal (todos os elementos se tocam, ou apenas alguns elementos se tocam). Algumas camadas encontram-se parcialmente dolomitizadas, silicificadas e com concreções de sílica, podendo estar associadas à dissolução. Os poros são vugulares de crescimento e intra-*shrubs*, com porosidade baixa a boa. Frequentemente esta fácies encontra-se intercalada com esferulititos, o que justifica a escolha de agrupamento entre estas fácies.

Esferulitito é uma rocha composta por partículas de formas esféricas ou subesféricas de contornos lisos ou lobados (esferulitos), de tamanho geralmente inferior a 2mm e que podem ocorrer de forma amalgamada ou isolados (Terra et al., 2010).

Nas amostras de testemunhos, foram identificadas três fácies cujas litologias correspondem a esferulititos e são diferentes em função do seu arcabouço: esferulitito maciço, esferulitito laminado e esferulitito caótico.

O esferulitito caótico é uma fácies que compreende os esferulititos com intensa modificação pela ação de fatores diagenéticos, causando uma trama “desorganizada” do arcabouço da rocha (Figura 7). Os esferulititos maciços e laminados ocorrem com esferulitos finos a grossos (diâmetros menores que 1mm) e baixa porosidade. O esferulitito maciço possui rara presença de lama entre os esferulitos e apresentam aspecto compacto e cristalino. Os esferulitos laminados, por sua vez, possuem presença de lama entre os grãos esferulíticos, com porções dolomitizadas e um aspecto recristalizado.

Por serem agrupamentos que representam as mesmas fácies, a caracterização petrofísica destes agrupamentos de fácies são bastante similares: os perfis dos conjuntos (Tabela 1) “Triplo combo”, “Sônico” e “Raios gamma + espectral” mostram valores muito similares para ambos agrupamentos de fácies, sendo os perfis de “Resistividade” e NPHI os únicos que demonstram valores significativamente distintos entre estas fácies, onde ESF-ARB possui baixa a moderada resistividade profunda (até 120 ohm), e a fácies ARB-ESF tem valores altos de resistividade profunda (> 150 ohm). Para o caso da porosidade NPHI, a predominância de porosidades maiores em fácies de calcários arbustiformes em comparação com fácies de esferulititos pode ser explicada pela

descrição petrográfica das rochas: os calcários arbustiformes podem ocorrer de forma vugular e com frequente dissolução, consistindo em uma porosidade intersticial gerada pela dissolução das argilas magnesianas singenéticas geradas durante a eodiagênese sob influência direta de composições de águas lacustres em resposta à elevada reatividade dos argilominerais magnesianos, que exercem fatores importantes de controle sobre a qualidade das rochas reservatório (Herlinger et al., 2017).

Uma característica deposicional em comum entre estes agrupamentos é que podem ocorrer mudstones intercalados entre os esferulititos e os calcários arbustiformes.

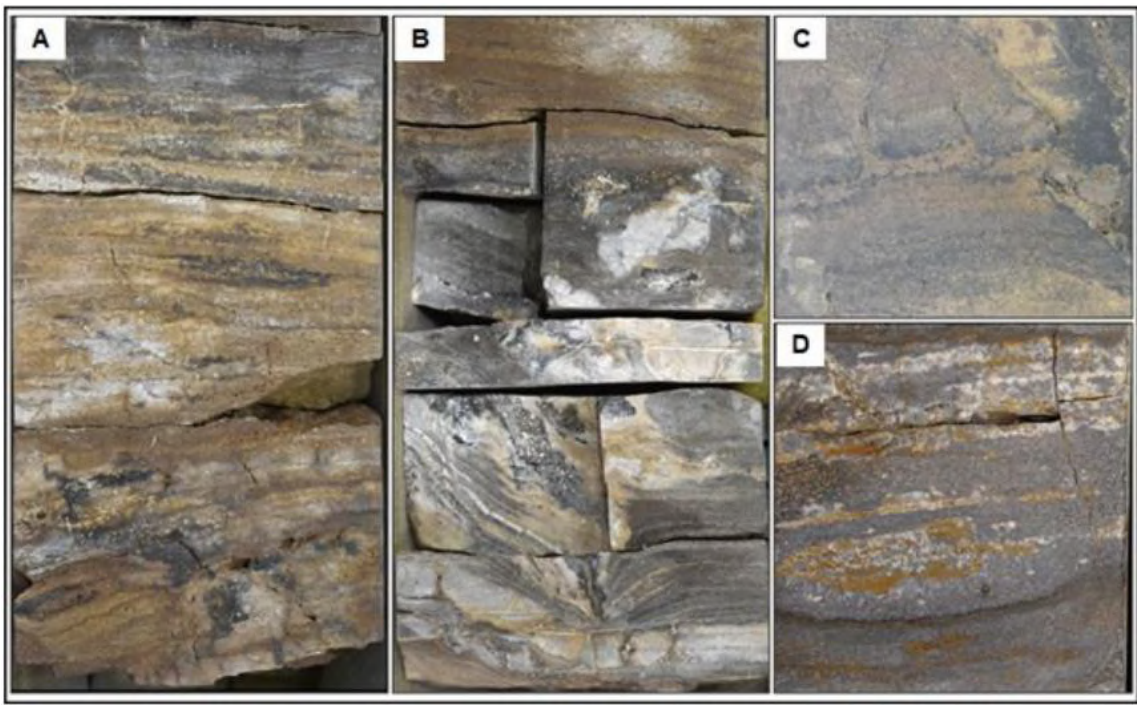

Figura 7: Esferulititos exibindo trama caótica em função de processos diagenéticos.

3.2.1.3 Calcário Arbustiforme-Arborescente (ARB-ARS) e Arbustiforme-Arborescente caótico (ARB-ARS-cao)

No agrupamento de fácies ARB-ARS ocorrem fácies de calcários arbustiformes maciços, arbustiformes vugulares, arborescentes e arborescentes vugulares. Calcários arborescentes são rochas carbonáticas que possuem os componentes internos que se organizam de forma ramificada e divergente e possuem comprimento maior que a largura (Terra et al., 2010).

Nos testemunhos descritos, a fácies calcário arborescente possui *shrubs* pequenos (2 mm – 5 mm) a médios (até 1,5 cm), com formas simples e complexas, onde os espaços entre os *shrubs* podem exibir poros de crescimento ou preenchimento por lama. A fácies arborescente vugular possui *shrubs* médios a grandes (> 5 mm), formas simples a complexas e empacotamento variável entre normal e aberto. Quando as fácies apresentam grau moderado a alto de dissolução, elas são classificadas como vugulares, e podem possuir finos cristais de dolomita romboédrica e outros sedimentos finos preenchendo os poros de crescimento que foram alargados pela dissolução (*vugs*), desenvolvendo boa porosidade aparente.

O agrupamento de fácies ARB-ARS-cao é designado para as fácies que sofreram intenso processo diagenético. O agrupamento é representado por calcários arborescentes caóticos, arbustiformes caóticos, esferulititos caóticos e silexitos. A intensa silicificação e/ou dolomitização das rochas ocorre associada a trama caótica do arcabouço destas fácies. O calcário arborescente caótico ocorre em camadas de espessura variável (poucos centímetros até 1m), ou como bolsões irregulares associados à fácies de calcário arbustiforme, calcário arborescente e calcário arborescente vugular. Nas fácies deste agrupamento, a porosidade é variável entre fechada a moderada, e é comum a ocorrência de *vugs* milimétricos. A fácies silexito também pertence a este agrupamento e ocorre como estratos tabulares, concreções centimétricas e laminações planares, frequentemente fraturados em padrão subvertical, com fraturas abertas ou fechadas preenchidas por calcita. Níveis mais espessos de silexito ocorrem com texturas brechadas, feições de deformação e esferulitos, laminações e *shrubs*, exibindo porosidade aparente baixa a regular.

Quantitativamente, estes agrupamentos de fácies são similares nos perfis petrofísicos de resistividade (moderada à alta, >100 ohm), de raios gamma (moderado a baixo, < 30ºAPI) e porosidade total (moderada, <12%), mas bastante distintos nos perfis litogeoquímicos de Ca e Si (Figura 8), onde as distribuições destes valores para cada agrupamento de fácies refletem os intensos processos diagenéticos das fácies “caóticas”, que sofreram dolomitização e silicificação.

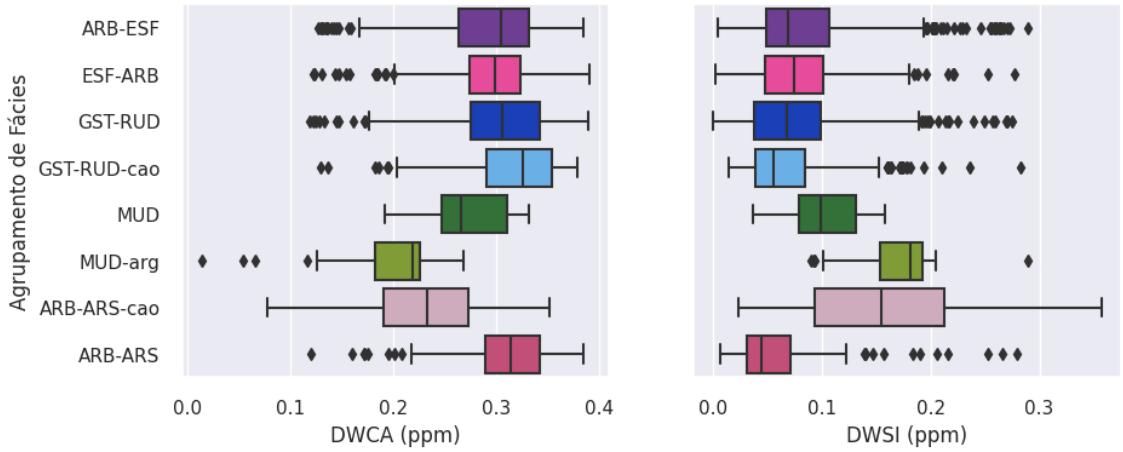

Figura 8: boxplot dos perfis litogeoquímicos de DWCA e DWSI para o agrupamento de fácies original

3.2.1.4 Grainstones-Rudstones (GST-RUD) e Grainstones-Rudstones caóticos (GST-RUD-cao)

Um "grainstone" é um termo proposto por Dunham (1962) para designar uma rocha carbonática suportada por grãos e com, no máximo, 5% de matriz. Um "rudstone", por sua vez, é um termo proposto por Embry & Klovan (1971) para designar uma rocha carbonática também suportada por grãos, sendo pelo menos 10% destes grãos maiores que 2mm. Os agrupamentos GST-RUD e GST-RUD-cao compreendem as fácies carbonáticas grainstone intraclástico e rudstone intraclástico, sendo o "GST-RUD-cao" formado por grainstones e rudstones caóticos, fácies cujo arcabouço possui trama modificada em função dos intensos processos diagenéticos que estas rochas sofreram.

Os grainstones intraclásticos foram descritos em escala de testemunho e são caracterizados pela ocorrência de intraclastos de retrabalhamentos das fácies anteriormente descritas, tais como os calcários arbustiformes, arborescentes e esferulititos. Os grãos possuem tamanho de areia fina a grossa (0.125 – 1mm) com grau de seleção moderado a baixo e arredondamento subanguloso a subarredondado. Ocorrem estruturas sedimentares que variam entre acamamento maciço e estratificação plano-paralela, podendo também apresentar estratificações cruzadas e intercalações com laminações milimétricas de sedimentos finos.

Os rudstones intraclásticos também são formados em função dos intraclastos dos retrabalhamentos das fácies dos agrupamentos anteriores. Os grãos possuem tamanho areia muito grossa a seixo, são mal selecionados e com arredondamento subanguloso a subarredondado. Frequentemente exibem estruturas sedimentares de estratificação plano-

paralela, podendo também ser cruzada, ou serem maciços. Os estratos costumam ocorrer com espessuras superiores a 20cm, frequentemente fraturados, com fraturas abertas e fechadas, e porosidade aparente boa, com poros intergranulares milimétricos.

As fácies "caóticas" correspondentes dos rudstones e grainstones estão agrupadas como GST-RUD-cao. Este agrupamento possui as fácies descritas anteriormente, porém caracterizadas por processos diagenéticos intensos, conferindo um arcabouço "desorganizado" a estas rochas, além de moderada à alta dolomitização (20-40%), e eventualmente com concreções de sílica e fraturas.

Em termos de atributos petrofísicos, estes agrupamentos são similares em caráter de argilosidade, com baixo à moderado Gamma Ray (<40ºAPI), resistividade profunda (>150 ohm) e litogeoquímica. São expressivamente diferentes nos perfis acústicos (VP, VS, DTCO e DTSM), em porosidade NPHI e densidade RHOB (Figura 9).

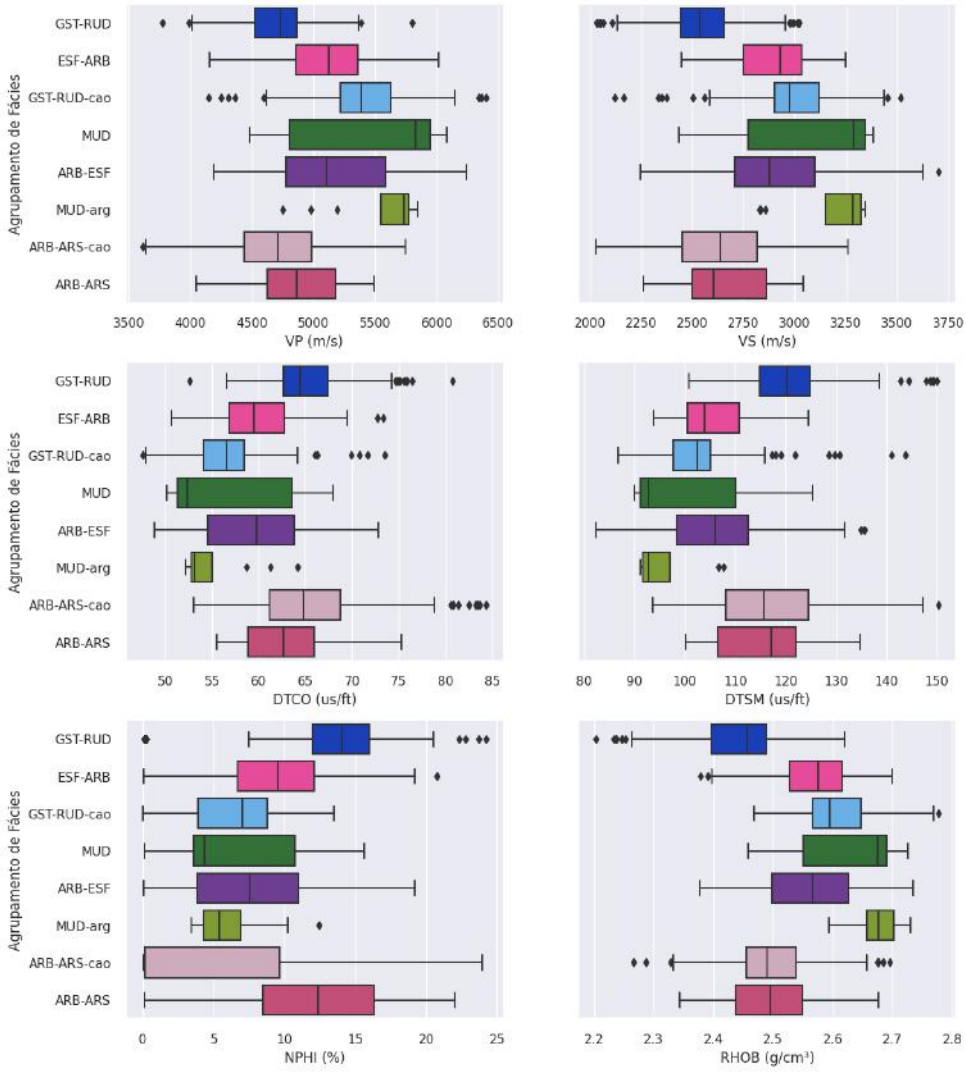

Figura 9: Distribuição em boxplot dos agrupamentos de fácies inicias para os perfis acústicos, NPHI e RHOB.

3.2.2 Reagrupamento de Fácies

A partir dos 8 agrupamentos de fácies iniciais do conjunto de dados original, algumas alterações entre estes agrupamentos foram realizadas para que a caracterização quantitativa destes agrupamentos seja o mais assertiva possível e contribua no processo de aprendizado do algoritmo de Machine-learning. Neste sentido, similaridades petrofísicas foram verificadas entre agrupamentos a partir de análises das distribuições estatísticas e estratigráficas. O resultado do reagrupamento final pode ser visualizado a partir dos 5 agrupamentos da Tabela 3, e as justificativas serão apresentadas a seguir.

Sigla do agrupamento	Fácies agrupadas em escala de poço
MUD-arg	Mudstone argiloso
ESF-ARB-ARS (MUD)	Calcário esferulítico-arbustiforme ou esferulítico-arborescente, podendo haver intercalações de mudstone
ARB-ARS-cao	Calcário arbustiforme-arborescente caótico
GST-RUD	Grainstones, rudstones
GST-RUD-cao	Grainstones e rudstones caóticos

Tabela 3: Reagrupamento final de fácies carbonáticas.

Com o novo reagrupamento, 5 classes foram estabelecidas a partir dos 8 originais. Para chegar a este número, os agrupamentos MUD, ESF-ARB, ARB-ESF e ARS-ARB tornaram-se um só: o agrupamento ESF-ARB-ARS (MUD). Este novo agrupamento faz sentido ao considerar simultaneamente atributos petrofísicos, petrográficos e estratigráficos.

A primeira decisão de agrupamento realizada foi a junção dos agrupamentos ESF-ARB e ARB-ESF por tratarem-se de agrupamentos com as mesmas fácies carbonáticas, sendo o fator de predominância entre esferulititos e calcários arbustiformes o critério para diferencia-los. Isto é traduzido em respostas petrofísicas também semelhantes, sendo os perfis de porosidade e resistividade os únicos que demonstram uma diferença significante entre estes agrupamentos.

Em seguida, um critério estratigráfico foi definido para os reagrupamentos restantes. Considerando os ciclotemas identificados por Wright e Barnett (2014) apresentados na Figura 4 e as descrições petrográficas nos testemunhos geológicos que demonstram a ocorrência de mudstones intercalados entre os calcários arbustiformes e

arborescentes, faz sentido agrupar as fácies calcário arborescente e mudstone tambérm, correspondentes aos agrupamentos ARB-ARS e MUD, gerando um novo agrupamento ESF-ARB-ARS (MUD). Desta maneira, cria-se um agrupamento que representa uma sequência deposicional completa.

A decisão de ter um agrupamento que define uma sequência deposicional pode favorecer o algoritmo de Machine-learning ao diminuir a dependência da profundidade. Na geologia existem padrões deposicionais que obedecem a ordens naturais, o que se traduz em uma dependência das variações litológicas ao longo da profundidade. Na Formação Barra Velha, por exemplo, pode-se dizer que existem sucessões estratigráficas que estabelecem uma continuidade espacial de fácies ao longo da profundidade, representando ciclos transgressivos-regressivos de deposição. Adicionar uma variável de "Profundidade" em um processo de treinamento de algoritmo de Machine-learning para identificar em que profundidades ocorrem quais litologias, no entanto, causa um grande viés, pois é uma variável que cresce monotonicamente, e por isso não é considerado uma boa prática (Jain et al., 2019). Ao invés disso, como as fácies destes agrupamentos ocorrem correlacionadas, opta-se por estabelecer um novo agrupamento com maior abrangência faciológica, mas que contém maior informação estratigráfica. Assim, espera-se diminuir a precisão preditiva em que cada fácies ocorre, mas aumenta-se a acurácia em que a sequência deposicional representada pelo agrupamento ESF-ARB-ARS (MUD) é reconhecida.

3.3 Machine-Learning Semi-Supervisionado

O termo em inglês "*Machine-learning*" representa um campo de estudo da estatística aplicada que se dedica em entender e desenvolver métodos estatísticos que "aprendem" informações a partir de dados para executar um determinado conjunto de tarefas e, assim, otimizar performances (Mitchell, 1997). É uma metodologia que é utilizada em diversas áreas do conhecimento científico, como a medicina, agricultura, e as geociências, sendo implementada principalmente quando se torna desafiador desenvolver um algoritmo explicitamente para realizar uma tarefa necessária.

Tradicionalmente, a abordagem de Machine-learning pode ser subdividida entre duas formas de execução: a abordagem supervisionada e a não-supervisionada. A forma de abordagem supervisionada implica na elaboração de um algoritmo que é construído a

partir de um modelo matemático utilizando um conjunto de dados que contém tanto as informações de entrada e as de saída, que representam as predições (Russell e Norvig, 2010). Este conjunto de dados é conhecido como o conjunto de treino, e representa os dados que serão utilizados para "ensinar" os padrões estatísticos de cada classe para o algoritmo utilizado. Os modelos matemáticos, por sua vez, realizam iterações com a finalidade de otimizar uma função objetiva para que o algoritmo de classificação supervisionada "aprenda" os padrões do conjunto de treino e os utilize para realizar predições a partir de novos dados de entrada.

A abordagem não-supervisionada utiliza apenas dados de entrada e se propõe a encontrar padrões nestes dados sem rotulá-los com uma classificação final. Uma prática comum do uso deste tipo de metodologia consiste em encontrar grupos ("*clusters*") estatísticos em dados não-rotulados multidimensionais. As formas de agrupar estes dados variam, podendo ser realizadas a partir de distâncias e densidades entre as amostras, por exemplo.

A abordagem utilizada neste trabalho é uma metodologia semi-supervisionada, que corresponde a uma forma intermediária entre as abordagens supervisionada e não-supervisiona (Jordan e Mitchell, 2015). Diferente destas abordagens, o Machine-learning semi-supervisionado faz uso de amostras rotuladas e não rotuladas simultaneamente durante a fase de treinamento do algoritmo. As amostras não-rotuladas fornecem informações importantes de como os dados estão distribuídos no espaço amostral, estabelecendo uma forma robusta de treinar um modelo com base em distribuições estatísticas (Mallapragada et al., 2009; Qi e Luo, 2020).

Um modelo esquemático simplificado adaptado de X. Ian et al. (2021) é demonstrado na Figura 10, que exibe como as amostras não-rotuladas são utilizadas para ajudar o modelo a realizar decisões. Neste exemplo simples, um problema de classificação binária em duas dimensões é apresentado, onde duas classes representadas por um circulo branco (classe 1) e um preto (classe 2) estão sendo classificadas a partir de um limite de classificação, representado pela linha vermelha pontilhada. As amostras cinzas representam as amostras não-rotuladas e a sua contribuição para ganho de informação no espaço amostral em (b). Desta forma, ao considerar a distribuição dos dados não-rotulados e um número limitado de amostras rotuladas ao mesmo tempo, é possível observar padrões estatísticos para o problema de classificação binária, e

estabelecer então uma forma mais razoável de calibrar os campos de classificação delimitados pela linha do limite de classificação em (c).

Neste trabalho, as profundidades com dados testemunhados e descritos são utilizadas como dados rotulados (com agrupamento de fácies estabelecido), enquanto os dados não-rotulados correspondem às informações das profundidades das amostras laterais e dos perfis petrofísicos nas profundidades não testemunhadas.

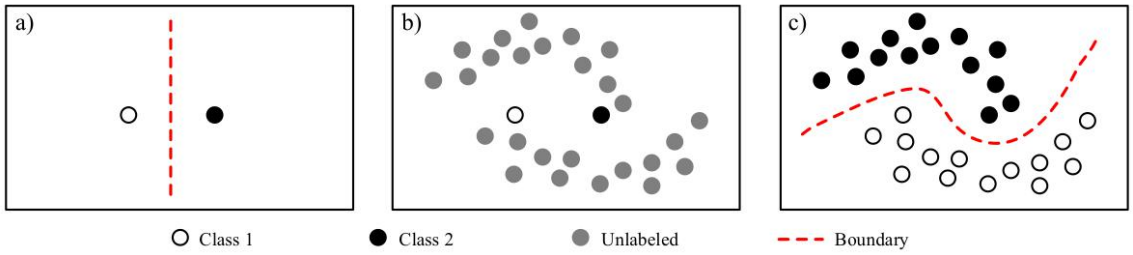

Figura 10: Exemplo de classificação semi-supervisionada a partir de um problema de classificação binária (X. Ian et al., 2021).

3.3.1 Abordagem Positive-Unlabeled Learning (PU-Learning)

Um caso especial de Machine-learning semi-supervisionado é aplicado neste trabalho, e trata-se da abordagem de "*Positive-Unlabeled Learning*" (PU-Learning). O PU-Learning é um esquema de classificação binária em que os dados são classificados entre "positivos" (1) e "desconhecidos" (0) (Scott e Blanchard, 2009; Claesen e De Smet, 2015). Existem três afirmações básicas sobre os dados que devem ser feitas ao se optar por uma abordagem de PU-learning (Bekker e Davis, 2020):

1. **Negatividade**: as amostras não-rotuladas pertencem à classe negativa (0). Isto permite que métodos convencionais de Machine-learning para classificação binária sejam utilizados.
2. **Separabilidade**: admite que os dados podem ser naturalmente separáveis. Isto significa que as amostras consideradas "positivas" podem ser distinguíveis das negativas a partir de um classificador apropriado.
3. **Suavidade**: amostras que estão mais próximas entre si no espaço amostral provavelmente pertencem à mesma classe.

Ao se tratar de um sistema de classificação binário, é necessário, portanto, que existam múltiplos classificadores, um correspondente para cada classe disponível no

conjunto de treino (padrão “um para um”). Este tipo de padrão permite compensar para classes que possuam menores frequências no conjunto de dados ao estabelecer um classificador específico que seja sensível a estas amostras (Sun et al., 2020). O agrupamento de fácies MUD-arg, por exemplo, possui poucas amostras no conjunto de dados, mas as suas características petrofísicas são suficientemente distintas das outras classes em termos de argilosidade, permitindo que um classificador específico para esta classe seja treinado para identifica-la, combatendo a generalização que um classificador treinado para todas as classes simultaneamente possuiria por consequência de um conjunto de dados desbalanceado.

Neste sentido, foram treinados cinco algoritmos, um para cada agrupamento de fácies. Primeiro, o conjunto de dados é dividido aleatoriamente em 70% para o conjunto de treinos, e 30% para o conjunto de testes, que será usado posteriormente para verificar as métricas estatísticas da classificação. As relações não-lineares entre os atributos de cada amostra e as classes correspondentes são estabelecidas durante a fase de treinamentos a partir de um classificador supervisionado.

Em seguida, para cada agrupamento de fácies, o fluxo de trabalho para a rotina de aprendizado desta abordagem classifica as amostras que correspondem à classe que o algoritmo pretende aprender como “positivas” (1), e as demais que correspondem à outras classes e amostras não definidas como “não-rotuladas” ou “negativas” (0). Por exemplo, supondo que o primeiro agrupamento que será treinado um algoritmo de aprendizado seja para o agrupamento “GST-RUD”, o código de execução irá tratar todas as amostras do conjunto de treinos que pertençam a esta classe como “1”, e as demais amostras, com agrupamento de fácies estabelecido ou não, serão chamadas de “0”. Em seguida, outros classificadores serão treinados para os demais agrupamentos, seguindo a mesma lógica de aprendizado.

Após todos os agrupamentos de fácies possuírem um classificador referente ao mesmo, é realizada a etapa de predição a partir do conjunto de testes. Para cada algoritmo de classificação treinado por agrupamento de fácies, é calculada a probabilidade de cada amostra rotulada do conjunto de testes pertencer a cada uma das classes. Para isso, essas amostras são utilizadas como entrada em cada um dos algoritmos treinados. O algoritmo, por sua vez, calcula a probabilidade de cada amostra pertencer ao agrupamento correspondente ao algoritmo. Ao final da utilização das amostras do conjunto de testes em todos os algoritmos de classificação de cada agrupamento, as probabilidades destas

amostras de pertenceram a cada um dos agrupamentos é comparada. Sabendo a probabilidade de cada amostra do conjunto de testes de pertencer a um agrupamento de fácies, a classe com a maior probabilidade é predita por amostra no conjunto de testes, e as métricas de avaliação como a precisão, a acurácia e a *f1-score* podem ser quantificados para posterior avaliação da performance geral do modelo.

3.4 Random-forest

As relações não-lineares entre os atributos de cada amostra e as classes correspondentes são estabelecidas durante a fase de treinamentos a partir de um classificador supervisionado. Qualquer classificador supervisionado que possa realizar a tarefa de classificação binária poderia ser utilizado. Neste trabalho, foi optado por utilizar o algoritmo "*Random-forest"* (florestas aleatórias) por se tratar de um algoritmo que evita o "*overfitting*" (sobreajuste) do modelo e ser de fácil implementação.

O algoritmo Random-forest é um método de aprendizado estatístico supervisionado utilizado tanto para tarefas de classificação quanto para regressão. Para tarefas de classificação, o Random-forest baseia-se na ideia de múltiplos classificadores de árvores de decisão geradas aleatoriamente para realizar predições em conjunto simultaneamente, estabelecendo um "comitê de decisão".

Árvores de decisão, por sua vez, são modelos preditivos que possuem uma estrutura que se assemelha a uma árvore, onde cada "folha" da árvore é uma classificação final e, cada galho (ou "ramo"), um filtro (ou decisão) que busca estabelecer um ganho de informação. A ideia do ganho de informação é determinar etapas que subdividem o conjunto de dados que se pretende classificar de forma que os mesmos se tornem os mais "puro" possíveis, isto é, sejam efetivos para determinar as diferenças entre as classes e, assim, realizar uma predição mais precisa.

Em geral, as árvores de decisão são classificadores fáceis de serem implementados e podem construir bons resultados, porém elas possuem um aspecto que as compromete em se tornarem uma abordagem ideal de aprendizado estatístico preditivo: a inacurácia (Hastie et al., 2008). Em síntese, as árvores de decisão não são flexíveis quando se trata de classificar novos dados, pois elas são muito adaptadas às características do conjunto de treinamento. Isto configura um viés estatístico que é designado o nome de "sobreajuste" ("*overfitting*").

O Random-forest combina a simplicidade das árvores de decisão com a flexibilidade, resultando em ganho de precisão. Neste algoritmo, cada árvore de decisão assume um voto para a classificação final, e a classe que obter mais votos é escolhida como predição do modelo (Fawagreh et al., 2014). A forma de decisão em que se estabelecem os nós das árvores é feita utilizando amostragens aleatórias, o que o torna menos propenso ao sobreajuste ao diminuir a correlação entre as árvores (Ghorbanzadeh e Blaschke, 2019).

Um exemplo simplificado de utilização do Random-forest para uma tarefa de classificação binária é exibido na Figura 11: em (a) é demonstrada a fase de treinamento de uma árvore de decisão onde os círculos representam os nós e as linhas representam os ramos (“galhos”) da árvore. Os círculos tracejados são as “raízes”, isto é, os nós “base” das árvores, e os nós em negrito correspondem às folhas, que dão a classificação final de uma amostra (neste caso, verdes ou vermelhas). Este processo de treinamento é executado por “N” árvores, o que leva a fase de classificação em (b). Nesta fase, ocorre um procedimento de votação majoritária entre todas as árvores simultaneamente, isto é, para cada nova amostra “X” do conjunto de testes, a amostra percorre cada árvore de decisão treinada, de cima para baixo, percorrendo os ramos e testando os valores de cada nó para chegar a uma decisão até chegar a uma folha, que estabelece uma classe. No final do processo, cada árvore vota para uma classe mais provável (o resultado da folha), e a que tiver mais escolhas entre as árvores é definida como predição final.

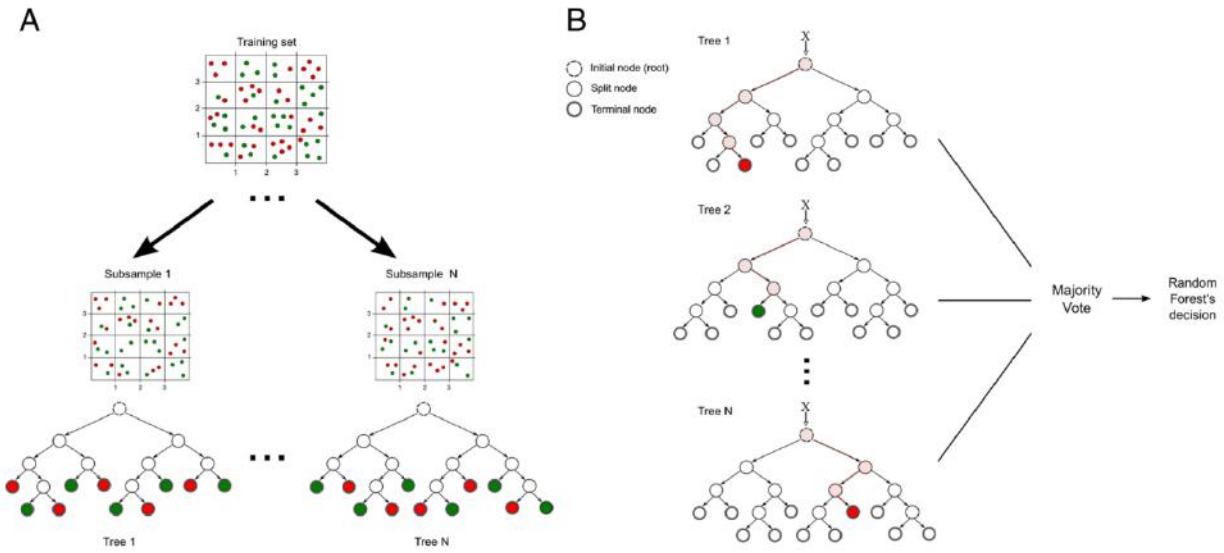

Figura 11: Exemplo de classificação binária a partir de árvores de decisão e florestas aleatórias. (Machado et al., 2015)

3.5 Métricas de Avaliação de Resultados

As principais métricas de avaliação de um modelo de Machine-learning são a precisão, a revocação, o F1-score e a acurácia. Para a estimação destes parâmetros, utiliza-se o conjunto de testes para realizar predições no algoritmo recém treinado, e compara-se os resultados obtidos com as classificações originais. Como os algoritmos foram treinados utilizando uma abordagem semi-supervisionada, os únicos dados utilizados para teste foram os que possuíam classificação a priori, que correspondem às amostras que foram interpretadas a partir de testemunhos geológicos.

A premissa básica para compreender os resultados dos modelos parte do entendimento do que são os quatro tipos de predições que se pode obter a partir do conjunto de testes: resultados positivos verdadeiros, positivos falsos, negativos falsos e negativos verdadeiros. Por exemplo, considerando um classificador binário da classe do agrupamento GST-RUD, supõe-se que a própria classe GST-RUD corresponde a um valor "positivo", e as demais classes um valor "negativo". Um resultado verdadeiro positivo, é quando o modelo prediz que a amostra é um GST-RUD, e a classe verdadeira é de fato um GST-RUD. De forma similar, um verdadeiro negativo é quando o modelo prevê corretamente a classe da amostra não se trata de um GST-RUD, e de fato a amostra pertencia a outro agrupamento de fácies (ex: MUD-arg). Seguindo a mesma lógica, um falso positivo corresponde a uma predição em que o modelo fez uma predição incorreta que diz que uma determinada amostra pertence a classe GST-RUD, e na verdade ela era outra classe, e um falso negativo corresponde a uma predição incorreta quando o modelo diz que a classe da amostra não é um GST-RUD, mas na verdade corresponde a esta classe. Um resumo destes resultados pode ser visualizado na Tabela 4.

		Valor verdadeiro	
		Positivo	**Negativo**
Valor Previsto	**Positivo**	**Verdadeiro Positivo**	**Falso Positivo**
	Negativo	**Falso Negativo**	**Verdadeiro Negativo**

Tabela 4: Verdadeiros positivos, verdadeiros negativos, falsos positivos e falsos negativos: parâmetros para métricas de avaliação dos modelos.

A partir das definições da Tabela 4, é possível estabelecer as relações das métricas de avaliação dentro do campo da recuperação de informação.

A precisão, também conhecida como "valor preditivo positivo", corresponde a fração de instâncias recuperadas que são relevantes. É definida na equação (1) como a razão entre os valores verdadeiros positivos (VP) pela soma dos valores verdadeiros positivos e falsos positivos (FP).

$$precisão = \frac{VP}{(VP + FP)} \quad (1)$$

Seguindo com o exemplo do classificador para a classe GST-RUD, a precisão para esta classe significa, então, a razão de quantos elementos selecionados são relevantes (verdadeiros positivos, ou seja, amostras GST-RUD que foram classificadas como tal) pela soma entre os verdadeiros positivos e os falsos positivos (amostras de outras classes consideradas falsamente como GST-RUD).

A equação da revocação, também conhecida como sensibilidade, é demonstrada na equação (2) e estabelece a razão das amostras relevantes (verdadeiros positivos) pela soma dos verdadeiros positivos e falsos negativos (FN).

$$revocação = \frac{VP}{(VP + FP)} \quad (2)$$

A revocação, portanto, quantifica a sensibilidade de quantas amostras relevantes (GST-RUD, para o caso do exemplo) foram classificadas corretamente no total.

A partir das definições das medidas de precisão e revocação, é possível estabelecer uma relação entre estas duas métricas de avaliação. O F1-score pode ser interpretado como uma média harmônica entre a precisão e a revocação, e a sua equação está definida em (3).

$$F1 = \frac{(precisão * revocação)}{(precisão + revocação)} \quad (3)$$

A derradeira métrica de avaliação que será calculada trata-se da acurácia. A acurácia representa a razão geral das predições que o modelo fez corretamente (VP + VN) e está representada em (4).

$$acurácia = \frac{(VP + VN)}{(VP + VN + FP + FN)} \quad (4)$$

3.6 Fluxo de Trabalho da Metodologia

A Figura 12 representa o fluxograma completo que foi estabelecido e aplicado neste trabalho. O fluxograma demonstra as etapas de pré-processamento, reagrupamento de fácies, aplicação dos algoritmos de *Machine-learning* e avaliação de resultados, que foram descritos anteriormente. Todas as etapas foram executas em código aberto e o fluxograma está documentado no repositório do *Github* referente a este trabalho: https://github.com/pvabreu7/TCC

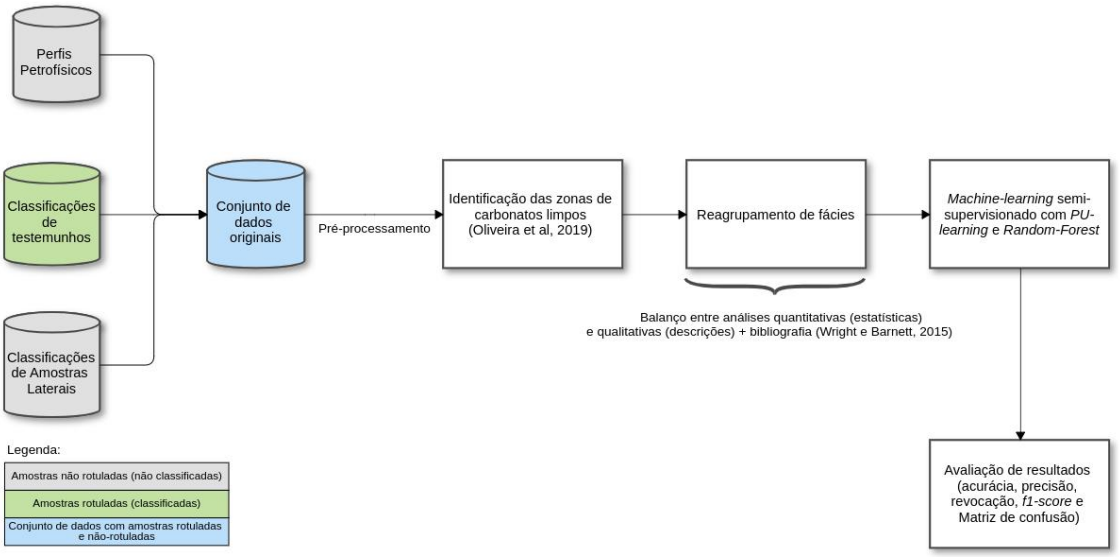

Figura 12: Fluxograma da metodologia aplicada neste trabalho.

4. RESULTADOS E DISCUSSÕES

Um algoritmo semi-supervisionado para cada classe correspondente a um agrupamento de fácies foi treinado a partir da abordagem *PU-learning* com um classificador *Random-forest* utilizando o conjunto de dados fornecidos pela Petrec e pela ANP e o reagrupamento de fácies estabelecido neste trabalho apresentado na Tabela 3. A quantidade de ocorrência de amostras para cada agrupamento de fácies pode ser observada no gráfico da Figura 13. Os resultados obtidos serão demonstrados e discutidos utilizando métricas gerais e individuais de precisão, revocação, acurácia e F1-score para cada agrupamento de fácies, além de exibição de gráficos das predições dos agrupamentos do conjunto de dados de teste e análise de matriz de confusão. Para efeito de avaliação da metodologia, estas métricas serão também comparadas com uma abordagem supervisionada realizada utilizando o agrupamento de dados original.

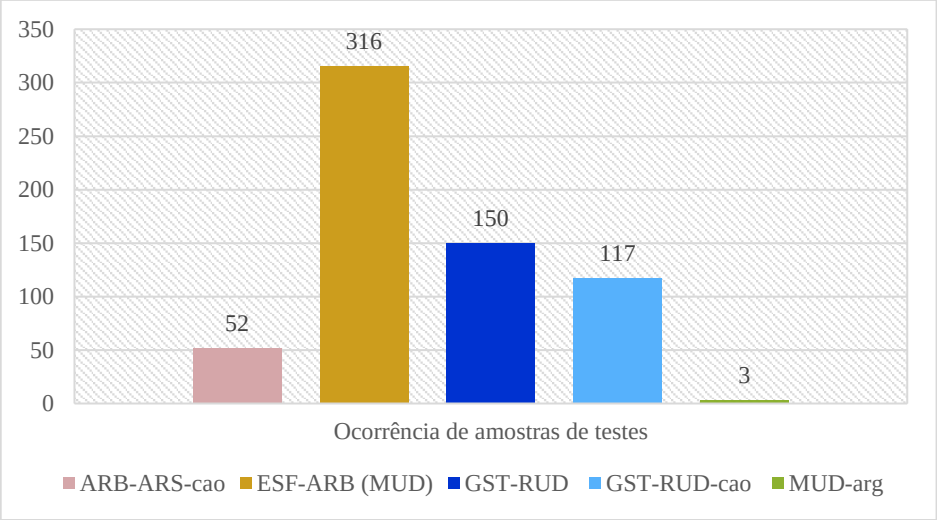

Figura 13: Quantidade de ocorrência de amostras para o conjunto de testes.

4.1 Precisão, revocação, acurácia, F1-score e matriz de confusão

Os dados do conjunto de treinos foram classificados para cada algoritmo treinado e as probabilidades de cada amostra de pertencerem a cada agrupamento de fácies foram calculadas. Para cada amostra, foi designada uma predição que corresponde ao agrupamento de fácies com a maior probabilidade de pertencimento.

A partir das predições realizadas, as métricas globais e individuais de avaliação da metodologia foram calculadas e podem ser visualizadas nos gráficos de barras da Figura 14 e da Figura 15.

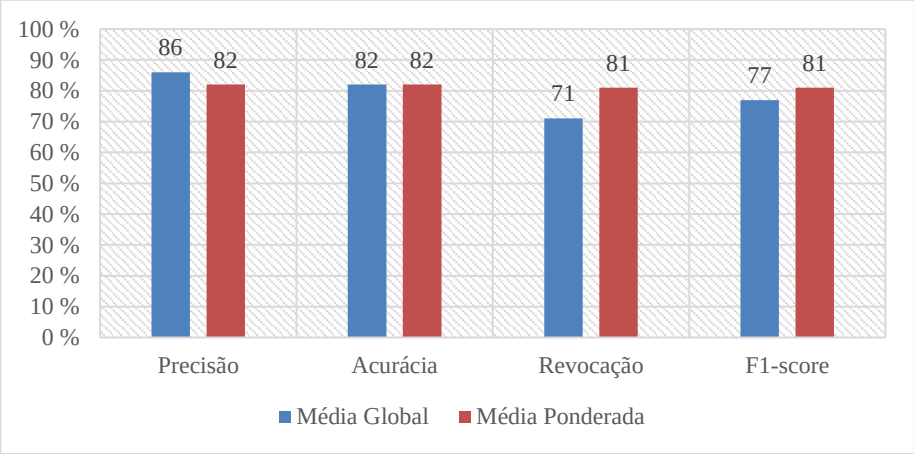

Figura 14: Médias globais e ponderadas para as métricas de avaliação do modelo.

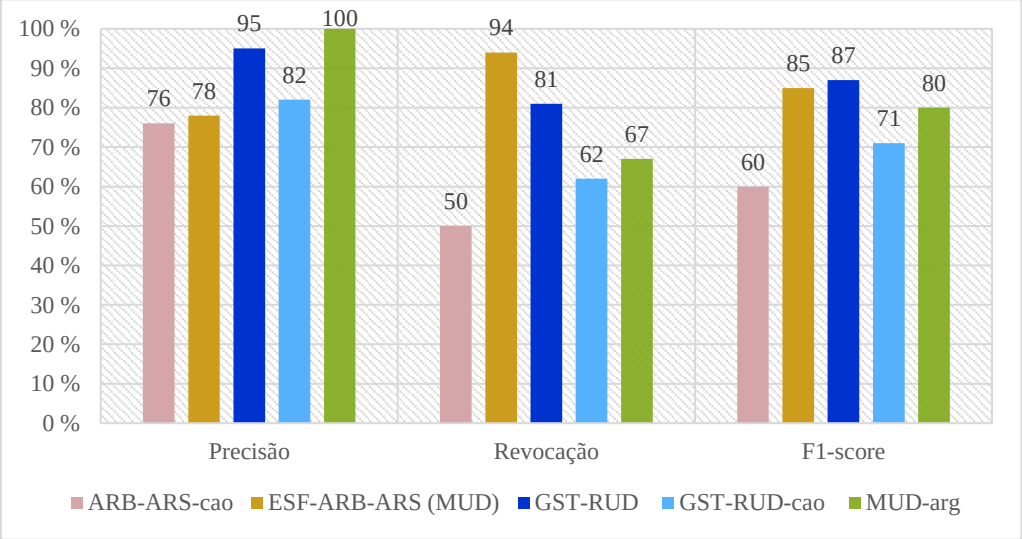

Figura 15: Métricas de avaliação individuais (para cada classe).

A avaliação dos resultados pelas métricas de avaliação do modelo demonstraram um elevado percentual de acerto na predição dos agrupamentos de fácies nas médias globais e ponderadas. O desbalanceamento da ocorrência de cada um dos agrupamentos de fácies no conjunto de dados impactou negativamente a revocação e o F1-score. Apesar disso, a revocação e o F1-score mantiveram resultados que são considerados elevados (81%).

Uma maneira de visualizar a performance da abordagem realizada é através de uma “matriz de confusão”. A matriz de confusão permite analisar a sensibilidade geral do

modelo em termos de cada classe individualmente. O visualizador mostra as predições em forma de matriz, em que cada linha representa um percentual de amostras de uma instância de classe do modelo, e cada coluna representa uma razão de amostras preditas para cada classe. A partir desta combinação entre as classificações reais das amostras do conjunto de testes e as suas respectivas predições realizadas pelo modelo proposto, é possível visualizar quais classes estão sendo mais “confundidas” pelo algoritmo, justificando este nome para esse método de visualização.

Na Figura 16 é exibida a matriz de confusão a partir dos resultados obtidos com os conjuntos de testes. Pode-se observar, portanto, que 40% das amostras relevantes do conjunto de testes pertencentes à classe ARB-ARS-cao foram confundidas com as amostras da classe ARB-ARS, e que 35% do agrupamento de fácies GST-RUD-cao e 33% das amostras de MUD-arg foram confundidos com ESF-ARB-ARS (MUD).

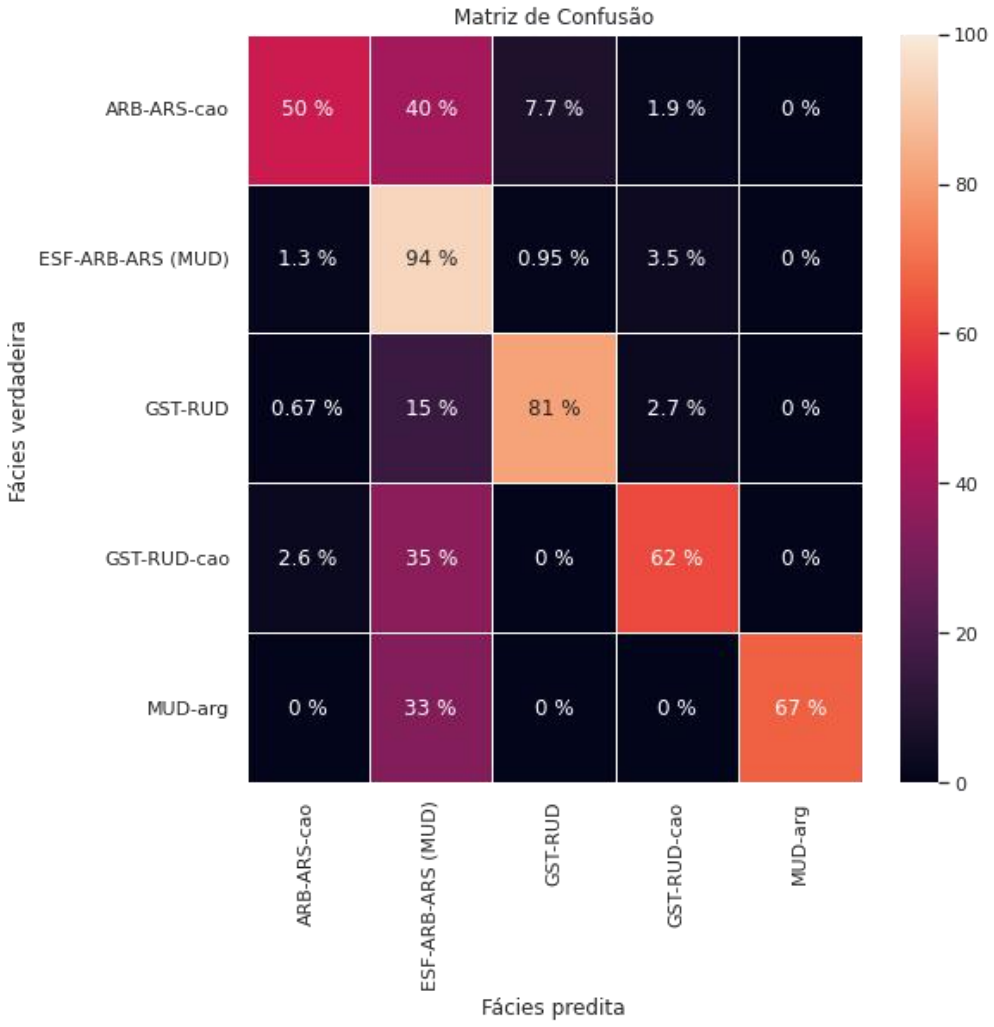

Figura 16: Matriz de confusão obtida pela abordagem semi-supervisionada com fácies reagrupadas.

A relação de confusão entre os agrupamentos de fácies ARB-ARS-cao e ARB-ARS pode ser justificada pela forma com que as fácies “caóticas” ocorrem nas descrições petrográficas baseadas nos testemunhos geológicos: elas formam “bolsões” dentro das rochas das fácies calcário arborescente e calcário arbustiforme. Essa relação diagenética entre as fácies destes agrupamentos compromete a precisão dos atributos petrofísicos que descrevem as fácies, que ocorrem diretamente associadas e, portanto, causam predições incorretas.

O agrupamento de fácies MUD-arg, por sua vez, apesar de ser bastante característico em função dos seus atributos petrofísicos no que diz respeito à argilosidade, ainda compartilha atributos semelhantes em outros perfis geofísicos com a fácies mudstone (do agrupamento MUD), presente no agrupamento de fácies ESF-ARB-ARS (MUD), o que leva à algumas predições incorretas. É necessário ressaltar, porém, que este agrupamento de fácies possui apenas 3 amostras no conjunto de testes, e isso torna as métricas de avaliação pouco representativas. O modelo demonstrou capacidade de identificar a fácies mudstone argiloso, no entanto, amostragens maiores que representem esta classe são necessárias para que a sua identificação seja plenamente validada.

Para fins de comparação entre uma abordagem tradicional supervisionada utilizando o algoritmo *Random-forest* com a metodologia realizada neste trabalho, um modelo único foi gerado para classificar todas os agrupamentos de fácies do conjunto de dados original de forma supervisionada. Os resultados podem ser observados no gráfico de barras da Figura 17.

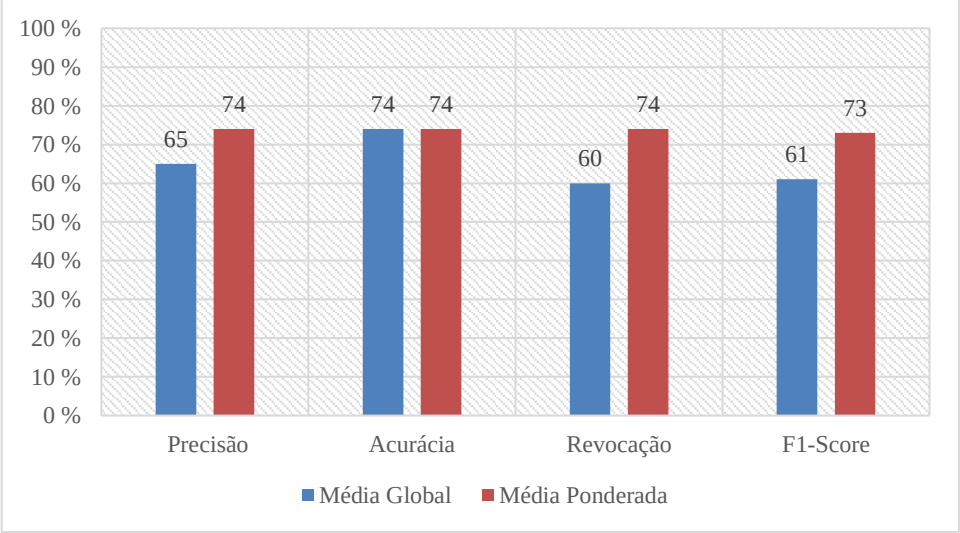

Figura 17: Resultados para abordagem supervisionada com o algoritmo *Random-forest*.

Observando os resultados da abordagem semi-supervisionada na Figura 13 e da abordagem supervisionada na Figura 16, se torna evidente um aumento expressivo nos resultados obtidos pela classificação semi-supervisionada. Algumas hipóteses podem ser levantadas para explicar estes números. Dentre elas, pode-se entender que um fluxo de trabalho semi-supervisionado consegue aproveitar melhor um conjunto de dados com frequência de classes desbalanceada ao permitir a otimização das informações rotuladas e não-rotuladas, como é o caso de diversos conjuntos de dados geocientíficos, principalmente no ramo da indústria de óleo e gás, onde dados com análises de detalhe como descrições petrográficas em rochas de testemunhos são mais escassos e de alto custo em relação à dados de perfis de poço, por exemplo.

Pode-se entender que, ao estabelecer uma única escala de detalhe para classificação de fácies (descrições de amostras de testemunhos geológicos), é possível padronizar as descrições faciológicas e adicionar ganho de informação a elas simultaneamente através das amostras não-rotuladas adicionadas a partir de amostras não-classificadas de respostas petrofísicas obtidas em perfilagem de poços e amostras laterais. Para o caso deste estudo, as amostras laterais não vieram classificadas pela ANP seguindo a mesma metodologia de descrição que os geocientistas da Petrec, também levantando a possibilidade destas amostras não corresponderem às mesmas fácies que as interpretadas pela Petrec e, por consequência, isto pode ter impactado negativamente os resultados do modelo supervisionado com os agrupamentos de fácies originais.

A criação de um agrupamento de fácies correspondente a uma sequência deposicional da Formação Barra Velha estabelecida através do grupo ESF-ARB-ARS (MUD) se mostrou uma ótima estratégia para identificação de fácies deposicionais, obtendo um f1-score com excelente confiabilidade demonstrada através de 85% de acerto. Esta estratégia, no entanto, remove a escala de resolução de identificação de fácies individuais ao agrupar diversas fácies que ocorrem em conjunto de maneira associada estratigraficamente. Apesar disso, a capacidade de possuir uma ferramenta com o caráter de identificação de uma importante sequência deposicional da Formação Barra Velha fornece uma interessante possibilidade de obter uma análise de suporte que auxilie em processos interpretativos com velocidade e confiabilidade.

4.2 Predições ao longo das profundidades

É possível afirmar que a identificação de agrupamentos de fácies que estejam associadas a sequências deposicionais permite que a problemática das ordens deposicionais em função da profundidade seja atenuada. Como descrito anteriormente, um tópico de pesquisa na área de inteligência computacional aplicada à indústria de óleo e gás consiste na construção de uma relação dos algoritmos de Machine-learning com relação a sequências estratigráficas. Naturalmente, as fácies ocorrrem associadas estratigraficamente em relação a profundidade, mas adicionar uma variável de "profundidade" no processo de aprendizado do algoritmo constituiria um forte viés ao mesmo, e o algoritmo poderia sofrer com sobreajuste ("*overfitting*"). Portanto, ao criar um agrupamento que englobe as fácies que ocorrem associadas, como o ESF-ARB-ARS (MUD), diminui-se a um certo grau a dependência da profundidade, mas também afeta a escala de predição, que se torna mais abrangente.

A questão da associação entre os agrupamentos de fácies, a boa relação de acerto na predição do conjunto de testes e a diminuição da escala de predição causada por agrupamentos que englobem sequências deposicionais podem ser visualizados nos gráficos que exibem as predições ao longo das profundidades para um poço do conjunto de dados exibido na Figura 18.

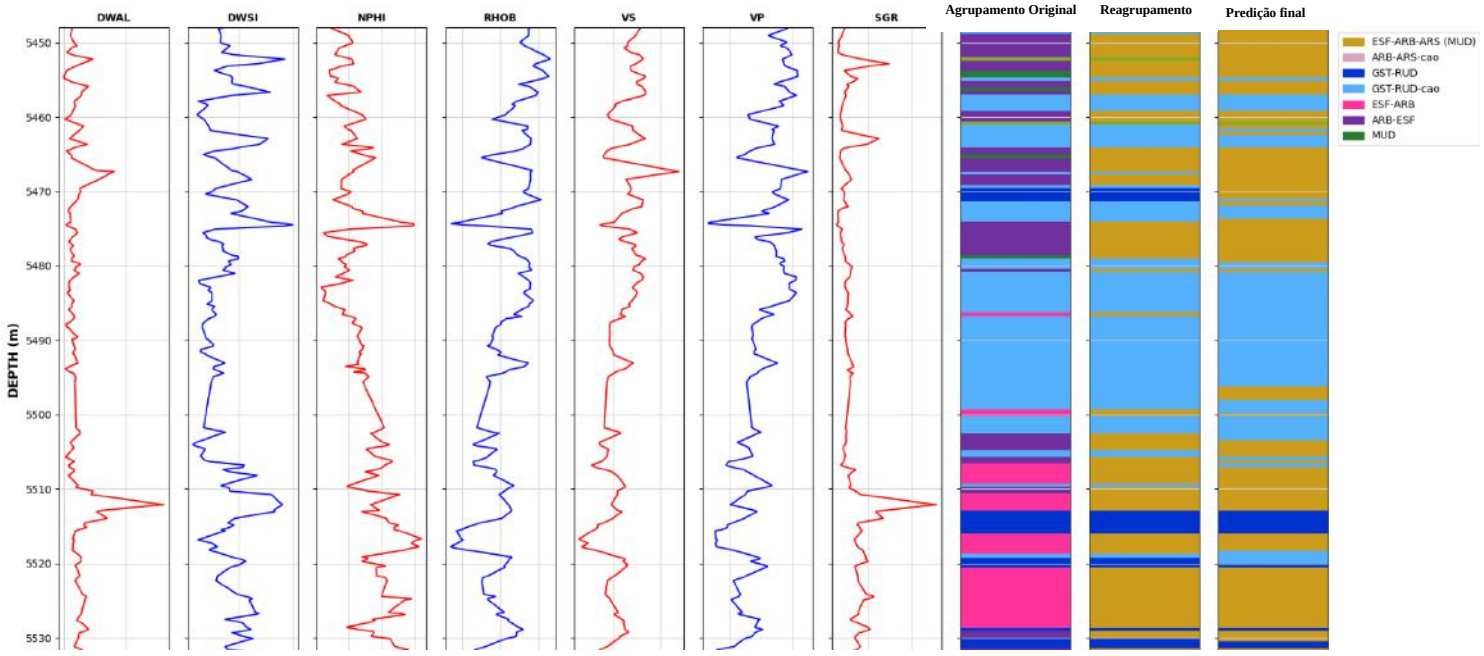

Figura 18: Predições ao longo de poço com o conjunto de testes

Nos gráficos do poço relativo à Figura 18 ocorrem todos os agrupamentos de fácies, originais e reagrupados, portanto, é possível observar alguns aspectos ligados ao modelo preditivo. O primeiro aspecto é o bom percentual de acertos das predições finais. O segundo é a diminuição do número de agrupamentos (diminuição da escala de predição de fácies) que vem acompanhado do reagrupamento de fácies. Por último, é a relação de ordem deposicional entre os agrupamentos de fácies. No agrupamento de fácies original, por exemplo, é possível verificar para este poço que mudstones ocorrem intercalados com o agrupamento ARB-ESF, que por sua vez ocorre associado à ESF-ARB e ARB-ARS. De forma semelhante, GST-RUD e GST-RUD-cao ocorrem também associados. Isto significa que, para um aprimoramento destes resultados, existe uma necessidade de estabelecer metodologias que adicionem conhecimentos da ordem de deposição de fácies ao longo das profundidades.

Propõe-se, neste sentido, duas metodologias que podem adicionar uma relação da ocorrência de fácies com a profundidade, de maneira a otimizar a resolução de predição de fácies em futuras publicações científicas que podem representar tópicos de pesquisa e aprimoração dos resultados deste trabalho.

A primeira delas foi utilizada no trabalho de Jain et al. (2019) e Wright e Barnett (2014), e trata-se do algoritmo de Cadeias de Markov. Cadeias de Markov são modelos estatísticos que são frequentemente utilizados para a detecção de padrões. Para isso, são calculadas probabilidades de transição de estado e associadas através de uma distribuição de probabilidade sobre possíveis resultados. Em Wright e Barnett (2014) as Cadeias de Markov foram utilizadas para evidenciar o reconhecimento do padrão de ocorrência de fácies e provar que as transições entre as fácies ao longo da profundidade não eram aleatórias (ou seja, são dependentes das fácies precedentes). Na publicação de Jain et al. (2019), uma metodologia de classificação é proposta em que as Cadeias de Markov são utilizadas após um processo de classificação supervisionada para estabelecer uma relação de probabilidade de ocorrência de cada fácies em função da profundidade e, assim, criar uma predição final que respeite estas associações deposicionais de fácies: “O uso de Cadeias Ocultas de Markov não só ajudou a solucionar a questão da dependência das profundidades, mas também ajudou a preservar a ordem natural das camadas em subsuperfície dado um conjunto de medições petrofísicas” (Jain et al., 2019).

A segunda abordagem que poderia ser testada para relacionar às profundidades com as fácies carbonáticas para predição com uso de Machine-learning seria a utilização

do Teorema de Bayes. O Teorema de Bayes consiste em uma atualização probabilística de ocorrência de um evento baseado em um conhecimento a priori cuja formulação está demonstrada na equação (5).

$$P(B) = \frac{P(A)\ x\ P(A)}{P(B)} \qquad (5)$$

Uma proposta seria associar probabilidades a priori de uma determinada fácies ocorrer em uma dada faixa de profundidade (P(B)) com a probabilidade a priori de uma amostra pertencer à uma determinada classe (P(A)). Se for possível calcular a probabilidade de que uma fácies pertença à um faixa de profundidade dado que ela tenha uma probabilidade à priori de pertencer a uma classe (P(B|A)), seria possível calcular a probabilidade a posterióri de que a amostra pertença à uma fácies dada que ela está condicionada à uma determinada profundidade (P(A|B)).

Por fim, trabalhos recentes foram publicados com o propósito de estabelecer padronizações nas descrições das rochas da Formação Barra Velha. A publicação de Borghi et al. (2022) estabelece uma linguagem em comum para a descrição das litologias carbonáticas da Formação Barra Velha que possa ser utilizada em amostras de diferentes escalas. Acredita-se que, por meio de uma descrição faciológica que se adeque a amostras de diferentes tipos (testemunhos, perfis geofísicos, amostras laterais), uma padronização entre as ocorrências faciológicas se torna mais fácil, e as assinaturas petrofísicas pertencentes à cada uma delas se tornem mais claras em termos de métodos estatísticos preditivos, fornecendo, assim, um tópico de pesquisa para aprimoramento futuro deste trabalho.

5. CONCLUSÕES

Uma metodologia de Machine-learning semi-supervisionado seguindo uma abordagem de PU-learning foi aplicada após o reagrupamento de fácies carbonáticas da Formação Barra Velha baseando-se em descrições quantitativas, qualitativas e critérios estratigráficos. Com isso, foi criada uma ferramenta preditiva capaz de identificar cinco agrupamentos de fácies carbonáticas com um bom grau de confiabilidade, revelando uma boa forma auxiliar de identificar sequências deposicionais, agrupamentos de fácies carbonáticas retrabalhadas, formadas *in-situ* e modificadas diagenéticamente.

O uso de algoritmos treinados de forma semi-supervisionada obteve bom êxito em lidar com um conjunto de dados com frequência de classes de forma desbalanceada, e isto fica evidente na comparação com os resultados obtidos com a aplicação de uma metodologia supervisionada tradicional, onde houve ganhos gerais de aproximadamente 10% nos resultados. A isso também pode-se atribuir a capacidade da metodologia de aproveitar mais dados de amostras classificadas e não-classificadas, unindo descrições técnicas com o ganho de informação de dados petrofísicos das amostras não-classificadas.

O reagrupamento de fácies estabeleceu classes que possuem significado geológico em termos deposicionais, que geraram predições com índices muito positivos de acertos preditivos, evidenciados pelas métricas de avaliação do modelo. Estes reagrupamentos, no entanto, diminuem a resolução da escala de predição ao englobarem diversas fácies carbonáticas de forma mais abrangente.

Nas discussões dos resultados, diferentes sugestões críticas foram feitas diante dos desafios inerentes a abordagens de Machine-learning relacionadas à predição de fácies. A aplicação de Cadeiras de Markov e o Teorema de Bayes foram citados como formas de aprimorar os resultados obtidos no sentido de poder incluir a relação da deposição de litologias ao longo das profundidades, assim estabelecendo uma maneira robusta do algoritmo aprender sobre a ordem natural da deposição das fácies que podem se tornar tópicos para futuras pesquisas.

Por fim, do ponto de vista do uso desta abordagem para a indústria de óleo e gás, a metodologia se mostrou conveniente por não exigir a mesma demanda de tempo e esforço técnico para interpretar todas as amostras em escala de testemunho e poço

simultaneamente com a finalidade de criar ferramentas de suporte para análise e tomadas de decisão.

Com base neste estudo, pode-se dizer que a tecnologia, quando aliada com o bom uso dos conhecimentos geocientíficos, mostra grande potencial para a otimização e a melhor utilização dos recursos naturais. Para tal aproveitamento, necessita-se, portanto, de um cuidado especial aos dados geocientíficos que, apesar de cada vez mais disponíveis em grande escala, demandam boa organização e padronização de resultados para o seu aproveitamento.

6. REFERÊNCIAS

Arienti, I. M., Souza, R.S., Viana, S., Cuglieri, M.A., Silva, R.P., Tonietto, S., Paula, I. De, Gil, J.A. (2018). Facies Association, depositional Systems, and Paleophysiographic Models of the Barra Velha Formation, Pre-Salt Sequence-Santos Basin, Brazil. In: AAPG ACE CONFERENCE, 2018, Salt Lake City. Abstract… Salt Lake City: AAPG Search and Discovery #2843310, 2p.

Aslanian, D., Moulin, M., Olivet, J.L., Unternehr, P., Matias, L., Bache, F., Rabineau, M., Nouzé, H., Klingelheofer, F., Contrucci, I., Labils,C. (2009). Brazilian and African passive margins of the Central Segment of the South Atlantic Ocean: Kinematic constraints: Tectonophysics, v. 468, 98-112.

Bekker J, Davis J. Learning from positive and unlabeled data: a survey. Machine-learning 2020;109(4):719–60.

Bize-Forest, N., Lima, L., Baines, V., Boyd, A., Abbots, F., & Barnett, A. (2018). Using machine-learning for depositional facies prediction in a complex carbonate reservoir. In SPWLA 59th Annual Logging Symposium. Society of Petrophysicists and Well-Log Analysts.

Borghi, L.; Correia, M.; Favoreto, Julia; Santos, Jeferson (2022). Definition a new common language: a multi-scale classification for the pre-salt carbonates of the Barra Velha Formation. Rio Oil & Gas Expo and Conference, 2022.

Buckley, J.P., Bosence, D., Elders. C. (2015). Tectonic setting and stratigraphic architecture of an Early Cretaceous lacustrine carbonate platform, Sugar Loaf High, Santos Basin, Brazil. In: Bosence, D. W. J., Gibbons, K. A., Le Heron, D. P., Morgan.

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, v. 38, n. 2, 29-46.

Chapelle O, Scholkopf B, Zien A. Semi-supervised learning. Cambridge, USA: MIT Press; 2006.

Claesen M, De Smet F, Suykens JAK, De Moor B. A robust ensemble approach to learn from positive and unlabeled data using SVM base models. Neurocomputing 2015;160:73–84.

De Oliveira, F. V. C. S., Gomes, R. T. M., & Silva, K. M. S. (2019). PS Log Features for the Characterization of Igneous Rocks in the Pre-Salt Area of Santos Basin, SE Brazil.

Dias, J. L. (2004). Tectônica, estratigrafia e sedimentação no Andar Aptiano da margem leste brasileira. Boletim Geociências Petrobras, Rio de Janeiro, v. 13, n. 1, 7-25, nov. 2004/ maio 2005.

Dunham, R. J. Classification of carbonate rocks according to depositional texture. (1962). In: Ham, W.E. (Ed.). Classification of carbonate rocks. Tulsa. American Association of Petroleum Geologists, Memeoir 1, p. 108-122.

Elkan C, Noko K. Learning classifiers from only positive and unlabeled data. In: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2008. P. 213-20.

Embry, A. F.; Klovan, J. E. A late Devonian reef tract on northeastern Banks Island, N.W.T. Bulletin of Canadian Petroleum Geology, v. 19, p. 730 - 781, 1971.

Fawagreh, K.; Gaber, M.; Elyan, E. Random forests: from early developments to recentadvancements. Systems Science & Control Engineering: An Open Access Journal, v. 2, n. 1, p. 602–609, 2014.

Folk, R. L. (1962). Spectral subdivision of limestones types. In Ham, W.E. (Ed.) Classification of carbonate rocks: Tulsa. American Association of Petroleum Geologists, Memoir 1, p. 62-85.

Garcia, Sávio Francis de Melo et al. Análise de volumes de sal em restauração estrutural: um exemplo na bacia de Santos. 2012. Revista Brasileira de Geociências, São Paulo, v. 42, n.2, p. 433-450, 2012.

Ghorbanzadeh, O.; Blaschke, T. Optimizing sample patches selection of CNN to improve the MIOU on landslide detection. GISTAM 2019 - Proceedings of the 5th International Conference on Geographical Information Systems Theory, Applications and Management, p. 33–40, 2019.

Gomes, P. O., Kilsdonk, B., Grow, T., Minken, J., Barragan, R. (2012). Tectonic evolution of the Outer High of Santos Basin, southern Sao Paulo Plateau, Brazil, and implications for hydrocarbon exploration. In: D. Gao, ed., Tectonics and sedimentation: Implications for petroleum systems: AAPG Memoir 100, 125-142.

Gomes, P.O. (2009). The Outer High of the Santos Basin, Southern São Paulo Plateau, Brazil: Pre-Salt Exploration Outbreak, Paleogeographic Setting, and Evolution of the Syn-Rift Structures. AAPG Search and Discovery, Article #10193.

Hastie, Trevor; Tibshirani, Robert; Friedman, Jerome (2008). The Elements of Statistical Learning (2nd ed.). Springer. ISBN 0-387-95284-5.

Herlinger Jr., R, Zambonato, E.E., De Ros, L. F. (2017). Influence of diagenesis on the quality of lower cretaceous pre-salt lacustrine carbonate reservoirs from nothern campos basin, offshore Brazil. Journal of Sed. Research, v. 87, 1285-1313.

Jain, Vikas, Wu, Po-Yen, Akkurt, Ridvan, Hodenfield, Brooke, Jiang, Tianmin, Maehara, Yuki, Sharma, Vipin, and Aria Abubakar. "Class-Based Machine Learning for Next-Generation Wellbore Data Processing and Interpretation." Paper presented at the SPWLA 60th Annual Logging Symposium, The Woodlands, Texas, USA, June 2019.

Jordan MI., Mitchell TM. Machine learning: Trends, perspectives, and prospects. Science 2015; 349(3245):255-60.

Machado, G..; Mendoza, M. R.; Corbellini, L. G. What variables are important in predicting bovine viral diarrhea virus? A random forest approach. Veterinary Research, v. 46, n. 1, 2015.

Mallapragada PK, Jin R, Jain AK, Liu Y. SemiBoost: Boosting for semi-supervised learning. IEEE Trans Pattern Anal Machine Intel 2009;31(11):2000–14.

Milani, E. J., Brandão, J. A. S. L., Zalán, P. V., Gamboa, L. A. P. (2000). Petróleo na margem continental brasileira: geologia, exploração, resultados e perspectivas. Brazilian Journal of Geophysics, v. 18, n. 3, 351-396.

Mitchell, Tom (1997). Machine Learning. New York: McGraw Hill. ISBN 0-07-042807-7. OCLC 36417892.

Momeni A, Rostami S, Hashemi S, Mosalman-Nejad H, Ahmadi A. Fracture and fluid flow paths analysis of an offshore carbonate reservoir using oil-based mud images and petrophysical logs. Mar Pet Geol 2019;109:349–60.

Moreira, J. L. P., Madeira, C. V., Gil, J. A. & Machado, M. A. P. (2007). Bacia de Santos. Boletim de Geociencas da Petrobras, 15, 531-549.

Muniz, M.C., & Bosence, D. W.J. (2015). Pre-salt microbialites from the Campos Basin (offshore Brazil): image log facies, faciesmodel and cyclicity in lacustrine carbonates. Geological Society, London, Special Publications.

Papaterra, G. E. Z. (2010). Pré-sal: conceituação geológica sobre uma nova fronteira exploratória no Brasil, 94p. Dissertação (Mestrado em Geologia) - UFRJ, Rio de Janeiro.

Qi GJ, Luo JB. Small data challenges in big data era: A survey of recent progress on unsupervised and semi-supervised methods. IEEE Trans Pattern Anal Machine Intel2020.

Ribeiro da Silva, Suzana; Figueiredo, Picanço; Coelho, Pedro Henrique; Borghi, Leonardo (2021). Evolução Tectônica Estratigráfica da Formação Barra Velha na área dos campos de Lapa e Sapinhoá, Bacia de Santos, Brasil. São Paulo, UNESP, Geociências, v. 40, n. 1, p. 55 – 69, 2021.

Riding, R. (2000). Microbial carbonates: the geological record of calcified bacterial-algal mats and biofilms. Sedimentology, v. 47, supplement 1, p. 179-214.

Russell, Stuart; Norvig, Peter (2003) [1995]. Artificial Intelligence: A Modern Approach (2nd ed.). Prentice Hall. ISBN 978-0137903955.

Saller, A., Rushton, S., Buambua, L., Inman, K., Mcneil, R., Dickson, J.A.D. (2016). Presalt stratigraphy and deposition systems in the Kwanza Basin, offshore Angola. AAPG Bulletin, v. 100, n. 7, 1135-1164.

Scott C, Blanchard G. Novelty detection: unlabeled data definitely help. In: Proceedings of the 12th International Conference on Artificial Intelligence and Statistics; 2009. p. 64–471

Souza, R. S., Arienti, L. M., Viana, S. M., Falcão, L. C., Cuglieri, M.A., Filho, R. P. S., Leite, C. O., Oliveira, V. C., Oliveira, D. M., Anjos, C., Amora, R., Carmo, I. D., Coelho, C. E. (2018). Petrology of the Hydrothermal and Evaporitic Continental Cretaceous (Aptian) Pre-Salt Carbonates and Associated Rocks, South Atlantic Santos Basin, Offshore Brazil. In: AAPG ACE CONFERENCE, 2018, Salt Lake City. Abstract... Salt Lake City: AAPG Search and Discovery #2835691, 2p.

Sun F, Fang F, Wang R, Wan Bo, Guo Q, Li H, et al. An impartial semi-supervised learning strategy for imbalanced classification on VHR images. Sensors 2020; 20 (22):6699.

Terra, G. J. S.; Spadini, A.R; França, A.B; Sombra, C.L; Zambonato, E.E.; Juschaks, L.C.S.; Arienti, L.M.; Erthal, M.M.; Blauth, M.; Franco, M.P.; Matsuda, N.S.; Silva, N.G.C.; Moretti Junior, P.A; D'Ávila, R.S.F.; Souza, R.S.; Tonietto, S.N.; Anjos, S.M.C.; Campinho, V.S.; Winter, W.R. (2010). Classificação de rochas carbonáticas aplicável às bacias sedimentares brasileiras. Boletim de Geociências da PETROBRAS, Rio de Janeiro, v. 18, n.1, p. 9-29.

Wright, V.P., & Barnett, A. (2017). Critically Evaluating the Current Depositional Models for the Pre-Salt Barra Velha Formation, Offshore Brazil. 51439, 1–40. N/A

Wright, P., Barnett, A. (2014). Ciclicity and carbonate-silicate gel interactions in Cretaceous Alkaline Lakes. AAPG Search and Discovery, Article #51011.

X. Lan, Zou. C., Kang Z., Wu X. (2019). Log facies identification in carbonate reservoirs using semi-supervised learning strategy. Fuel, v. 302, 121-145, 2021.