UNIVERSIDADE FEDERAL DO MARANHÃO
CENTRO DE CIÊNCIAS EXATAS E TECNOLÓGICAS
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA

LUIZ EUGÊNIO HOFFMANN LOPES

CLASSIFICAÇÃO DE SEDIMENTOS DA MARGEM EQUATORIAL BRASILEIRA UTILIZANDO ALGORITMOS DE APRENDIZADO DE MÁQUINA

São Luís - MA
2024

LUIZ EUGÊNIO HOFFMANN LOPES

CLASSIFICAÇÃO DE SEDIMENTOS DA MARGEM EQUATORIAL BRASILEIRA UTILIZANDO ALGORITMOS DE APRENDIZADO DE MÁQUINA

Dissertação apresentada ao Programa de Pós-Graduação em Engenharia Elétrica da Universidade Federal do Maranhão, como parte dos requisitos necessários para obtenção do grau de Mestre em Engenharia Elétrica.

Orientador: Prof. Dr. Allan Kardec Duailibe Barros Filho
Coorientador: Prof. Dra. Marta de Oliveira Barreiros

São Luís - MA

2024

Ficha gerada por meio do SIGAA/Biblioteca com dados fornecidos pelo(a) autor(a).
Diretoria Integrada de Bibliotecas/UFMA

Lopes, Luiz Eugênio Hoffmann.
Classificação de sedimentos da Margem Equatorial Brasileira utilizando algoritmos de aprendizado de máquina / Luiz Eugênio Hoffmann Lopes. - 2024.
67 f.

Orientador(a): Allan Kardec Duailibe Barros Filho.
Dissertação (Mestrado) - Programa de Pós-graduação em Engenharia Elétrica/ccet, Universidade Federal do Maranhão, São Luís / MA, 2024.

1. Aprendizado de Máquina. 2. Classificação. 3. Margem Equatorial Brasileira. 4. Sedimentos Marinhos. I. Filho, Allan Kardec Duailibe Barros. II. Título.

LUIZ EUGÊNIO HOFFMANN LOPES

CLASSIFICAÇÃO DE SEDIMENTOS DA MARGEM EQUATORIAL BRASILEIRA UTILIZANDO ALGORITMOS DE APRENDIZADO DE MÁQUINA

Dissertação apresentada ao Programa de Pós-Graduação em Engenharia Elétrica da Universidade Federal do Maranhão, como parte dos requisitos necessários para obtenção do grau de Mestre em Engenharia Elétrica.

Aprovado em:

BANCA EXAMINADORA

Prof. Dr. Allan Kardec Duailibe Barros Filho (Orientador)
Universidade Federal do Maranhão (UFMA)

Prof. Dra. Marta de Oliveira Barreiros (Coorientador)
Universidade do Estado do Pará (UEPA)

Prof. Dr. Ewaldo Éder Santana (1º Membro)
Universidade Federal do Maranhão (UFMA)

Prof. Dra. Priscila Lima Rocha (2º Membro)
Instituto Federal de Educação, Ciência e Tecnologia do Maranhão (IFMA)

Este trabalho é dedicado aos meus pais,
irmãos, cunhados e a minha querida esposa Andrea.

AGRADECIMENTOS

Agradeço a Deus pela vida.

Aos meus pais Luiz Fernandes Lopes e Maria do Carmo Hoffmann Lopes, por todo o apoio dado durante a minha vida e, sempre incentivando na busca por conhecimento e novas leituras.

Aos meus irmãos Mariângela e Luiz Fernando me apoiando nas decisões que tomo.

À minha querida esposa, meu amor e melhor amiga Andrea Carolina Aires Hoffmann, com quem divido minha vida, frustrações e conquistas. É meu porto seguro.

Ao Prof. Dr. Allan Kardec Duailibe Barros Filho, pela confiança, oportunidade e não ter desistido de mim.

À Profa. Dra. Marta de Oliveira Barreiros e ao Prof. Me Diego de Oliveira Dantas pelo direcionamento inicial e o incentivo em seguir nesta linha de pesquisa, que acreditando no meu potencial, me ajudaram a realizar todas as etapas desta pesquisa.

Ao Programa de Pós-Graduação em Engenharia de Eletricidade da UFMA.

Ao apoio financeiro da Agência Nacional do Petróleo, Gás Natural e Biocombustíveis (ANP) e da Financiadora de Estudos e Projetos (FINEP), por meio do Programa de Recursos Humanos da ANP para o Setor Petróleo e Gás - PRH/ANP (PRH 54.1 - UFMA).

À Agência Nacional do Petróleo, Gás Natural e Biocombustíveis (ANP) e à TGS pelo fornecimento dos dados de sísmica multicanal.

A todos que contribuíram diretamente ou indiretamente na produção deste trabalho.

RESUMO

Recentemente, importantes discussões a respeito da ocorrência de extensos recifes carbonáticos e do potencial de descobertas de hidrocarbonetos nas bacias do Pará-Maranhão e Foz do Amazonas, elevaram a importância do estudo da composição faciológica da plataforma continental. No entanto, amostrar sedimentos marinhos é uma operação de alto custo e muitas vezes demorada a depender da distância da costa e da profundidade da amostragem, sendo necessário o deslocamento da equipe e dos equipamentos através de embarcações de pesquisa especializadas. Tendo isto em vista, o objetivo principal deste trabalho foi integrar informações sedimentológicas de amostras de fundo e atributos sísmicos a partir de algoritmos de aprendizado de máquina para assim otimizar o mapeamento das fácies na plataforma continental da Margem Equatorial Brasileira. Os dados de sedimentos foram adquiridos do Banco Nacional de Dados Oceanográficos (BNDO), sendo classificados em Lama, Areia e Cascalho para o tamanho do grão e em Litoclástico (<30%), Litobioclástico (<50%), Biolitoclástico (<70%) e Bioclástico (>70%) para o teor de carbonato de cálcio. Os atributos sísmicos utilizados foram: Coerências, Amplitude, Intensidade, Fase Instantânea e Frequência Instantânea. Foram desenvolvidos cinco algoritmos de aprendizado de máquina para relacionar os tipos de sedimento e teor de carbonatos aos atributos sísmicos. Os resultados de cada algoritmo são Support Vector Machine: Accuracy 82,21%, Precision 84,11%, Recall 83,11% e F1_Score 83,66%. K-Nearest Neighbors: Accuracy 73,22%, Precision 78,47%, Recall 73,22% e F1_Score 75,74%. Decision Trees: Accuracy 76,63%, Precision 79,33%, Recall 76,63% e F1_Score 77,92%. Multilayer Perceptron: Accuracy 83,99%, Precision 84,56%, Recall 83,99% e F1_Score 84,27%. Adaptive Boosting: Accuracy 74,38%, Precision 72,82%, Recall 74,38% e F1_Score 73,35%. O melhor resultado obtido foi do algoritmo Multilayer Perceptron com 83,99% de acurácia. Com esses resultados preliminares, observa-se que é possível criar um modelo de aprendizado de máquina para classificar automaticamente os sedimentos marinhos, viabilizando e otimizando pesquisas futuras sobre as fácies carbonáticas e recifes mesofóticos na Margem Equatorial Brasileira.

Palavras-chave: Aprendizado de Máquina. Classificação. Margem Equatorial Brasileira. Sedimentos Marinhos.

ABSTRACT

Recently, important discussions regarding the occurrence of extensive carbonate reefs and the potential for hydrocarbon discoveries in the Pará-Maranhão and Foz do Amazonas basins have raised the importance of studying the faciological composition of the continental shelf. However, sampling marine sediments is a high-cost and often time-consuming operation, depending on the distance from the coast and the sampling depth, requiring the transportation of staff and equipment via specialized research vessels. With this in mind, the main objective of this work was to integrate sedimentological information from bottom samples and seismic attributes using machine learning algorithms to optimize the mapping of facies on the continental shelf of the Brazilian Equatorial Margin. Sediment data were acquired from the National Oceanographic Data Bank (BNDO), being classified into Mud, Sand and Gravel for grain size and into Lithoclastic (<30%), Lithobioclastic (<50%), Biolithoclastic (< 70%) and Bioclastic (>70%) for calcium carbonate content. The seismic attributes used were coherence, Amplitude, Intensity, Instantaneous Phase and Instantaneous Frequency. Five machine learning algorithms were developed to relate sediment types and carbonate content to seismic attributes. The results of each algorithm are Support Vector Machine: Accuracy 82.21%, Precision 84.11%, Recall 83.11% and F1_Score 83.66%. K-Nearest Neighbors: Accuracy 73.22%, Precision 78.47%, Recall 73.22% and F1_Score 75.74%. Decision Trees: Accuracy 76.63%, Precision 79.33%, Recall 76.63% and F1_Score 77.92%. Multilayer Perceptron: Accuracy 83.99%, Precision 84.56%, Recall 83.99% and F1_Score 84.27%. Adaptive Boosting: Accuracy 74.38%, Precision 72.82%, Recall 74.38% and F1_Score 73.35%. The best result obtained was from the Multilayer Perceptron algorithm with 83.99% accuracy. With these preliminary results, it is possible to create a machine learning model to automatically classify marine sediments, enabling and optimizing future research on carbonate facies and mesophotic reefs in the Brazilian Equatorial Margin.

Keywords: Machine Learning. Classification. Brazilian Equatorial Margin. Sediments Marines.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

BNDO	Banco Nacional de Dados Oceanográficos
BDCA	Banco de Dados de Caracterização Ambiental
BAMPETRO	Banco de Dados da Indústria do Petróleo
CNN	Convolutional Neural Networks
CSV	Comma-Separated Values
FN	False Negative
FP	False Positive
KNN	K-Nearest Neighbors
ML	Machine Learning
MLP	Multilayer Perceptron
RNA	Redes Neurais Artificias
SIG	Sistema de Informações Geográficas
SVM	Support Vector Machine
TN	True Negative
TP	True Positive
ZEE	Zona Econômica Exclusiva

SUMÁRIO

1 INTRODUÇÃO

A crescente demanda por energia impulsiona a Margem Equatorial Brasileira para o centro das atenções como uma promissora fronteira para a extração de recursos não renováveis, destacando-se o petróleo e o gás natural. Esta região ganha relevância devido às semelhanças geológicas que compartilha com a costa ocidental da África, já reconhecida mundialmente por suas vastas reservas petrolíferas. A analogia entre as margens Brasileira e Africana abre caminho para a possibilidade de condições petrolíferas similares no Brasil, como corroborado por descobertas recentes de reservatórios nas bacias de Barreirinhas e Pará-Maranhão (PELLEGRINI e RIBEIRO, 2018).

A Margem Equatorial já abriga cerca de 18 campos produtores de petróleo, concentrados principalmente nas Bacias Potiguar e Ceará, consolidando-se como uma área estratégica para a exploração de recursos energéticos (ANP, 2020). A extensão aproximada de 9.650 km de litoral oferece condições favoráveis para a implementação de fazendas eólicas offshore no Brasil. Contudo, as políticas de regulação para licenciamento, monitoramento e proteção ambiental dessas instalações continuam em fase de discussão e desenvolvimento (GONZÁLEZ et al., 2020).

Nesse contexto, as medidas regulatórias, o planejamento estratégico e a ocupação do espaço marítimo demandam um refinamento no reconhecimento do leito marinho, especialmente nas áreas destinadas à exploração de petróleo e gás. Dada a extensa região costeira do Brasil e a margem continental adjacente que, por vezes, ultrapassa as 200 milhas (aproximadamente 322 quilômetros) náuticas da Zona Econômica Exclusiva (ZEE).

O tipo de sedimento do fundo do mar é um parâmetro ambiental importante, sua distribuição tem significado científico e prático para a pesquisa em ciência marinha, construção e exploração de recursos marinhos. Com o desenvolvimento e utilização em larga escala dos recursos marinhos, é importante o desenvolvimento de tecnologias para detectar e compreender os tipos de distribuição dos sedimentos do fundo do mar. A aquisição sem erros de informações sobre a distribuição de sedimentos no fundo do mar desempenha um papel importante na construção de bancos de dados geográficos marinhos. A amostragem extensiva dos sedimentos marinhos não se mostra eficiente ou operacionalmente viável para o reconhecimento ideal das fácies sedimentares da plataforma continental.

Para superar essa limitação, este trabalho utiliza dados públicos de sedimento e sísmica multicanal 2D disponíveis na região da Margem Equatorial Brasileira. O objetivo é prever a classe de sedimento por meio de técnicas de aprendizado de máquina. O tipo de sedimento no fundo do mar é um parâmetro ambiental crucial, com relevância científica e prática para a pesquisa em ciências marinhas, construção e exploração de recursos marinhos.

1.1 Objetivos

A seguir é apresentado o objetivo geral deste trabalho e também os objetivos específicos.

1.1.1 Objetivo Geral

Classificar os sedimentos marinhos da região da Margem Equatorial Brasileira, mediante dados sísmicos, dados sedimentológicos e algoritmos de aprendizado de máquina.

1.1.2 Objetivos Específicos

- Fazer uma revisão na literatura sobre a classificação de sedimentos do mar;
- Organizar uma base de dados sobre aspectos de sedimentos do mar da Margem Equatorial Brasileira;
- Implementar algoritmos de aprendizado de máquina para classificação dos padrões de sedimentos do mar, usando *Support Vector Machine* (SVM), *K-Nearest Neighbors* (KNN), *Adaptive Boosting*, *Decision Trees* e *Multilayer Perceptron* (MLP);
- Avaliar a eficiência dos algoritmos propostos através das métricas: *Accuracy*, *Precision*, *Recall* e *F1-Score*;
- Comparar diferentes algoritmos de aprendizado de máquina para a classificação de sedimentos da Margem Equatorial Brasileira.

1.2 Trabalhos Relacionados

A classificação de sedimentos marinhos desempenha um papel crucial em várias aplicações, incluindo defesa/naval, ambiental e indústria marítima (MAYER et al., 2018). Diversos métodos foram empregados para a classificação e mapeamento automático de sedimentos, abrangendo desde técnicas tradicionais até abordagens de aprendizado de máquina (IERODIACONOU et al., 2011; DARTNELL; GARDNER, 2004). Métodos supervisionados, como *Decision Trees*, *Random Forest*, Support Vector Machines (SVM), *Rede Neural Artificial* (RNA) e o algoritmo *AdaBoost*, têm sido amplamente utilizados (LUCIEER et al., 2013; LI; TRAN; SIWABESSY, 2016; HASAN; IERODIACONOU; MONK, 2012; FRANCKE; LóPEZ-TARAZóN; SCHRöDER, 2008). Entre eles, o *Random Forest* tem se destacado em estudos comparativos (LI; TRAN; SIWABESSY, 2016; DIESING et al., 2020).

Enquanto métodos tradicionais são comuns, técnicas de aprendizado profundo, como as *Convolutional Neural Networks* (CNNs), ganharam destaque (CUI et al., 2021;

QIN et al., 2021). Especificamente, a arquitetura U-Net tem sido aplicada com sucesso em diversas tarefas, desde a segmentação de habitats marinhos até a classificação de leitos rochosos (AROSIO et al., 2023).

A automação da criação de mapas de sedimentos ainda enfrenta desafios, apesar dos avanços (DIESING et al., 2020). Estudos recentes indicam a necessidade de métodos alternativos (BROWN et al., 2011), e embora várias abordagens tenham sido testadas, a comparação sistemática de diferentes métodos ainda é limitada (HASAN; IERODIACONOU; MONK, 2012; DIESING et al., 2020).

Ou seja, a classificação de sedimentos marinhos envolve um conjunto diversificado de métodos, desde abordagens tradicionais até técnicas de aprendizado de máquina, sinalizando um campo de pesquisa em constante evolução na busca por métodos eficazes e automatizados.

1.3 Estrutura do Trabalho

Este Trabalho apresenta uma comparação de diferentes algoritmos de aprendizado de máquina na classificação de sedimentos marinhos. Portanto, no capítulo 2 aborda os tópicos da Margem Equatorial Brasileira, Sedimentos Marinhos, Sísmica Multicanal 2D e os algoritmos: *Multilayer Perceptron*, *Decision Trees*, *K-Nearest Neighbors*, *AdaBoost* e *Support Vector Machine*.

No capítulo 3 são apresentadas a metodologia utilizada neste trabalho, bem como a base de dados, como foi feito o pré-processamento dos dados após a aquisição, o desenvolvimento dos algoritmos e as métricas de avaliação de desempenho.

No capítulo 4 e 5 são apresentados os resultados e discussões sobre a análise dados nos experimentos, que mostram os algoritmos mais eficientes na classificação dos sedimentos da região da Margem Equatorial Brasileira.

Por fim, no capítulo 5 são abordadas as considerações finais deste trabalho e as propostas para trabalhos futuros.

2 FUNDAMENTAÇÃO TEÓRICA

Este capítulo traz uma revisão utilizada como base para a elaboração do trabalho. Explicando Margem Equatorial Brasileira, Sedimentos Marinhos, Sísmica Multicanal 2D e Aprendizado de Máquina.

2.1 Margem Equatorial Brasileira

A Margem Equatorial Brasileira, situada ao longo do litoral nordeste do país, é uma área geologicamente rica e estratégica, cuja formação remonta a eventos fundamentais na história geológica da Terra. Esta margem representa uma extensão da plataforma continental brasileira, abrangendo desde o sul da Argentina até o extremo nordeste da costa brasileira. Sua formação está intrinsecamente ligada aos mecanismos de distensão litosférica que ocorreram a partir do Mesozoico (MILANI et al., 2000).

O processo de distensão litosférica resultou na ruptura do antigo supercontinente Gondwana, levando à separação definitiva das placas Africana e Sul-Americana. Esse fenômeno acompanhou a formação do Oceano Atlântico Sul, delineando a configuração atual da Margem Equatorial Brasileira. A região é marcada por uma área predominantemente distensiva, estendendo-se desde o sul da Argentina, seguindo até o extremo nordeste da costa brasileira. Além disso, um segmento com natureza transformante, conhecido como Atlântico Equatorial, destaca-se nesse contexto, onde processos distensionais ainda ocorrem atualmente (MILANI et al., 2000).

As bacias sedimentares que compõem a Margem Equatorial desempenham um papel crucial na exploração de recursos naturais, especialmente petróleo e gás. Originadas no Neo-Triássico e Cretáceo, essas bacias apresentam características distintas em comparação com outras regiões da plataforma continental brasileira. O mecanismo de cisalhamento dextrogiro é responsável pela ruptura crustal na região, resultando em um padrão de falhas oblíquas subverticais que controlam o rifteamento, evoluindo para grandes zonas de fraturas oceânicas, como a zona de fratura de São Paulo e Romanche (MILANI et al., 2000).

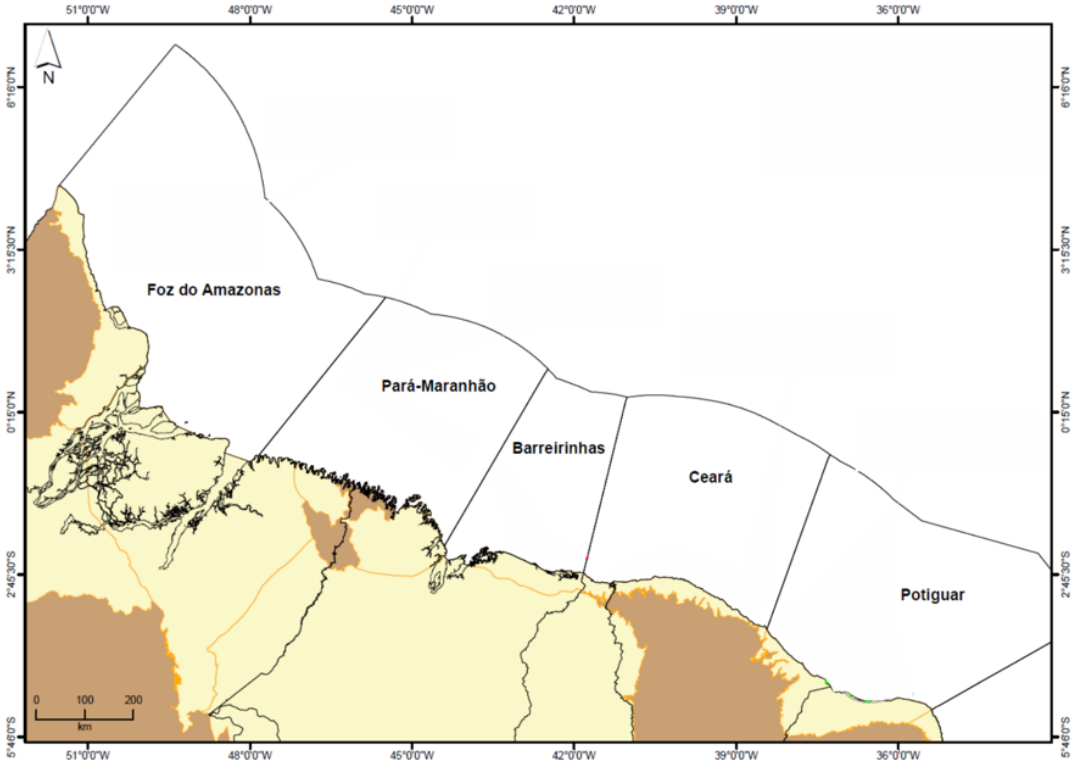
Uma particularidade digna de nota é a presença de um trecho ao norte da Foz do Amazonas, que faz parte do contexto distensivo do Oceano Atlântico Central, sendo mais antigo que o restante da Margem Equatorial Brasileira (MILANI et al., 2000). As bacias ao longo desta margem compartilham fases tectonosedimentares evolutivas comuns, caracterizadas por uma fase rifte, uma fase transicional e uma fase marinha aberta.

Em termos de importância, a Margem Equatorial Brasileira destaca-se como uma área estratégica para a exploração de recursos energéticos e minerais. Suas bacias sedimentares, com suas características geológicas únicas, são alvos de intensas pesquisas e

atividades de exploração, contribuindo significativamente para o desenvolvimento econômico do país. O entendimento detalhado da geologia e dos processos geodinâmicos ao longo desta margem é crucial para orientar as práticas sustentáveis de exploração e preservação ambiental na região.

A margem é composta pelas bacias da Foz do Amazonas, Pará-Maranhão, Barreirinhas, Ceará e Potiguar, destacadas na Figura 1. Estas bacias possuem juntas, aproximadamente 576.997 Km^2 (MILANI et al., 2000).

Figura 1 – Margem Equatorial Brasileira.

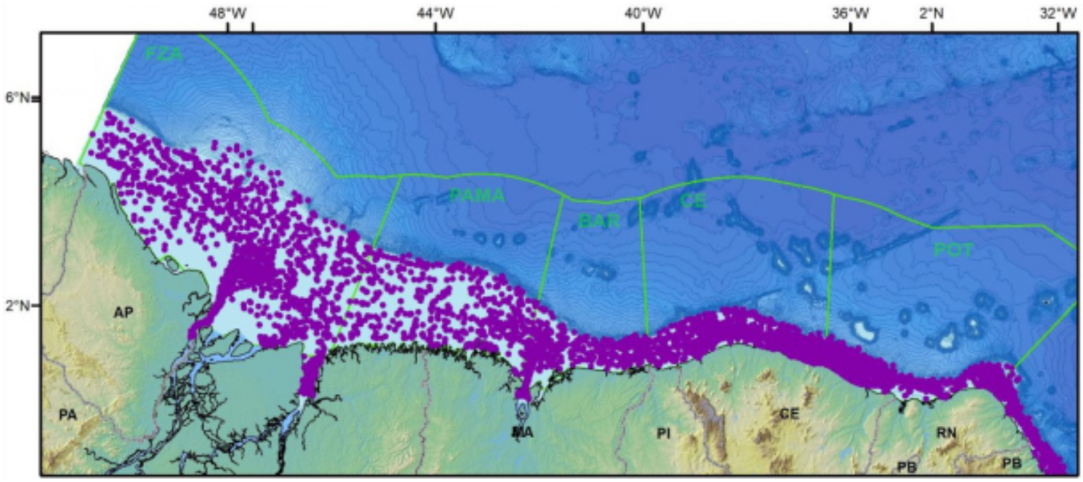


Fonte: adaptada (ANP, 2022).

2.2 Sedimentos Marinhos

Os dados sobre sedimentos superficiais resultam da integração de diversas fontes, fazendo parte do conjunto do Banco Nacional de Dados para a Indústria do Petróleo (BAMPETRO) e do Banco de Dados de Caracterização Ambiental (BDCA) (FILHO et al., 2022). Essas bases de dados reúnem uma ampla variedade de amostras de sedimentos superficiais coletadas pelo BNDO, bem como contribuições de trabalhos acadêmicos e projetos de monitoramento ambiental. O período de aquisição de dados do BAMPETRO abrange 1956 a 2004, totalizando 48 anos e 30.633 pontos de sedimento amostrados ao longo da margem brasileira (conforme Figura 2).

Figura 2 – Coordenadas dos sedimentos.



Fonte: Retirado de (FILHO et al., 2022).

Os dados podem ser categorizados em dois tipos principais: descrição visual do sedimento (tença) e resultados de análises laboratoriais. As tenças são obtidas por meio da observação visual das amostras de fundo no momento da coleta, com o objetivo de classificar a natureza do leito marinho para garantir a segurança da navegação. Essas classificações seguem as diretrizes estabelecidas pelo Manual de Hidrografia da Organização Hidrográfica Internacional (OHI) (International Hydrographic Organization, 2005).

Por outro lado, os dados provenientes do BDCA correspondem exclusivamente a amostras que foram analisadas em laboratório. Essas amostras são oriundas das atividades de monitoramento ambiental vinculadas a diversos empreendimentos de exploração e produção de petróleo.

Para este trabalho, foram utilizados ao todo 5.515 dados de sedimentos superficiais provenientes do Banco Nacional de Dados Oceanográficos (BNDO) e do Banco de Dados da Indústria do Petróleo (BAMPETRO), estes correspondentes à plataforma continental da Margem Equatorial Brasileira, que inclui cinco bacias sedimentares marginais distintas: Foz do Amazonas (FZA), Pará-Maranhão (PAMA), Barreirinhas (BAR), Ceará (CE) e Potiguar (POT).

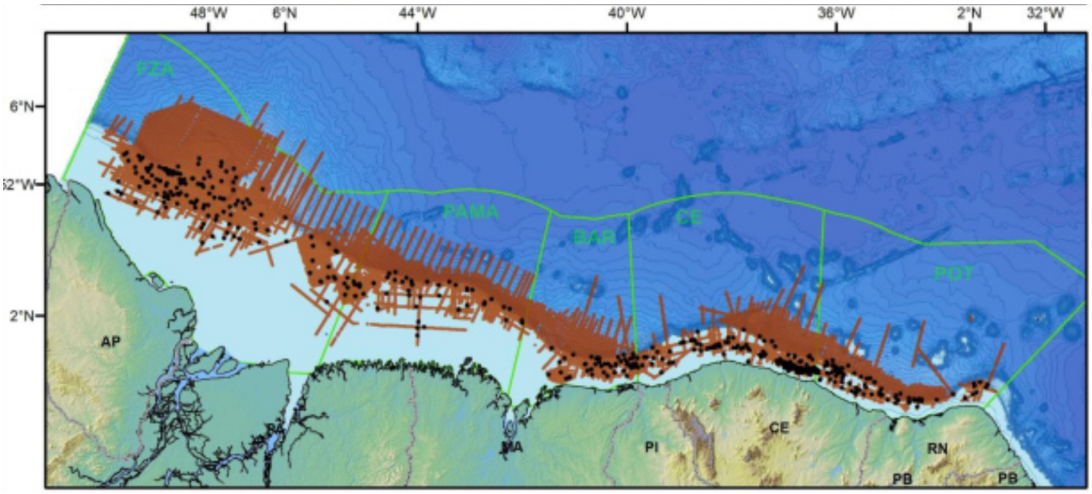
2.3 Sísmica Multicanal 2D

Os dados sísmicos são obtidos por meio de técnicas de sísmica, uma abordagem amplamente empregada na exploração de recursos subterrâneos, como petróleo e gás. As linhas sísmicas representam registros das ondas sísmicas emitidas e refletidas pelas camadas subterrâneas, fornecendo informações valiosas sobre a estrutura geológica. São considerados atributos sísmicos, tais como amplitude, intensidade de reflexão, fase, frequência e coerência, que são parâmetros essenciais para caracterizar as propriedades das camadas geológicas.

Esses dados desempenham um papel fundamental em análises e tomadas de decisão em atividades de exploração e produção de recursos naturais.

Os dados de atributos sísmicos utilizados neste estudo consistem em linhas sísmicas multicanais 2D públicas (Figura 3), processadas no software PaleoscanTM. Na Bacia da Margem Equatorial Brasileira, foram identificadas 384 linhas sísmicas por meio de pesquisa no banco de dados. Para cada uma dessas linhas, os primeiros 500 milissegundos foram interpretados manualmente para extração de cinco atributos sísmicos: amplitude, intensidade de reflexão, fase, frequência e coerência (FILHO et al., 2022).

Figura 3 – Linhas sismicas.



Fonte: Retirado de (FILHO et al., 2022).

Essas informações foram então organizadas para criar uma matriz com cinco atributos. O Sistema de Informações Geográficas (SIG) extraiu os valores desses atributos da trajetória das linhas sísmicas até os pontos das estações de sedimento, que estavam localizados a uma distância mínima de 500 metros.

2.4 Aprendizado de Máquina

Machine Learning (ML) representa uma evolução constante dos algoritmos computacionais, concebida para imitar a inteligência humana ao aprender com estímulos ambientais (NAQA; MURPHY, 2015). Originando-se na década de 1960 como parte da inteligência artificial, focada na identificação de padrões em bancos de dados, a ML evoluiu para um campo independente na década de 1990, estabelecendo interseções significativas com a estatística (IZBICKI; SANTOS, 2020). Hoje, sua aplicação abrange diversas áreas, como reconhecimento de padrões, visão computacional, engenharia espacial, finanças, entretenimento e biologia computacional (NAQA; MURPHY, 2015).

Os algoritmos de ML são categorizados quanto à forma de aprendizado supervisionado e não supervisionado. No aprendizado supervisionado, o algoritmo faz predições

com base nos rótulos (valores da variável resposta) observados, enquanto no não supervisionado, o aprendizado ocorre mesmo na ausência desses rótulos (IZBICKI; SANTOS, 2020). A aprendizagem supervisionada aborda dois problemas principais: classificação, que gera uma categorização qualitativa dos dados a partir do processamento dos dados de entrada; e regressão, que produz um valor numérico como resultado (SILVA; VELASQUEZ, 2020). Entre os algoritmos de ML notáveis, destacam-se, *Multilayer Perceptron*, *Decision Trees*, *K-Nearest Neighbors*, e *Support Vector Machine*.

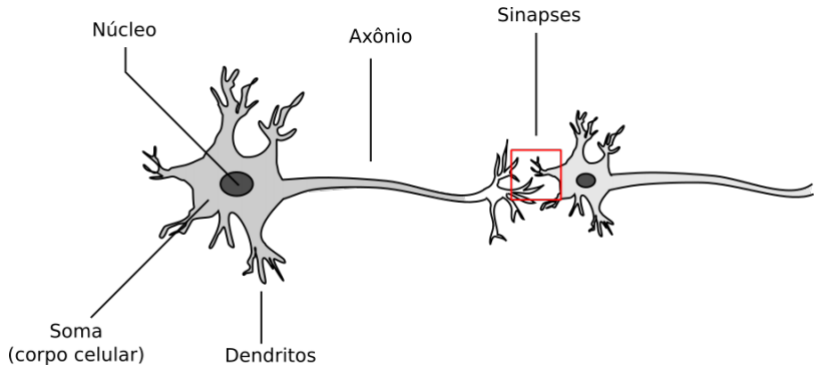
2.4.1 Redes Neurais Artificiais

Redes Neurais Artificiais (RNA) são sistemas computacionais com inspiração no sistema nervoso de seres vivos e que adquire conhecimento através da experiência. Os neurônios artificiais são interligados por sinapses artificiais representadas por vetores de peso sináptico (HAYKIN, 2007). As RNAs solucionam diversos problemas que necessitam de alto processamento de dados. Os principais problemas que as RNAs podem resolver são classificação de padrões, aproximação de funções, reconhecimento de fala, visão computacional e processamento de texto.

2.4.1.1 Neurônio

O sistema nervoso possui um conjunto complexo de células, por exemplo, os neurônios biológicos responsáveis pelo funcionamento e comportamento do corpo e da mente. O neurônio biológico é uma célula dividida em 3 partes distintas: **os dendritos** responsáveis por receber informações, **o corpo celular** responsável por armazenar as informações e **o axônio** responsável por enviar as informações para os próximos neurônios (AZEVEDO, 2016). A figura 4 abaixo ilustra um neurônio biológico.

Figura 4 – Estrutura básica de um neurônio biológico.

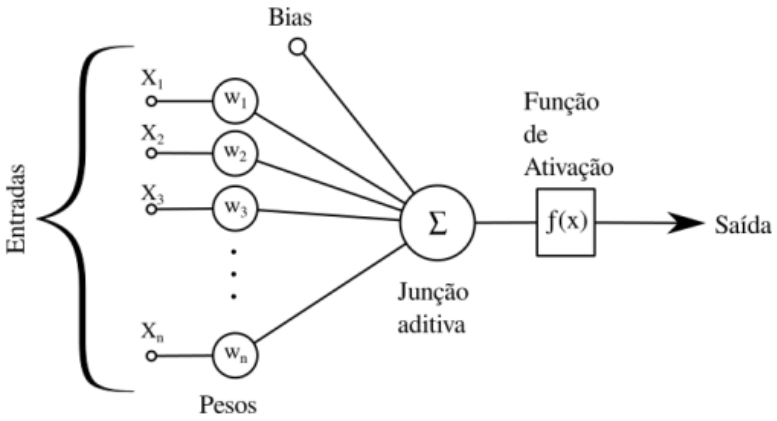


Fonte: Retirado de (AZEVEDO, 2016).

De forma, análoga, o neurônio artificial (apresentado na figura 5) possui funcionamento semelhante ao neurônio biológico, sendo a unidade de processamento fundamental

para a operação de uma Rede Neural Artificial com inspiração em um neurônio biológico, é composto pelos seguintes elementos: **conjunto de entradas**, **somador** e **função de ativação** (HAYKIN, 2007).

Figura 5 – Neurônio Artificial.



Fonte: Retirado de (AZEVEDO, 2016).

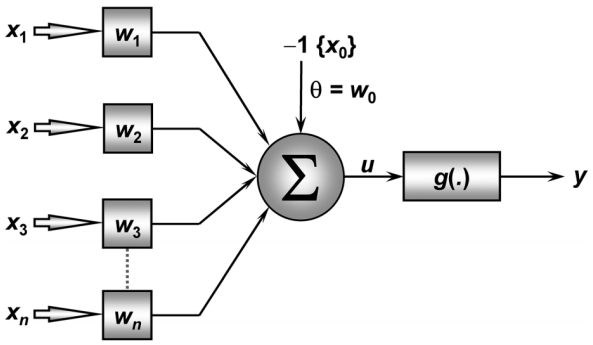
2.4.1.2 *Perceptron*

Segundo Silva, Spatti e Flauzino (2016) o *Perceptron* é a configuração de RNA mais simples, pois é constituído por apenas uma camada e tem apenas um neurônio nela. Tinha como aplicação inicialmente identificar padrões geométricos.

Só consegue classificar padrões com apenas duas classes linearmente separáveis, exemplo figura 7, sendo utilizada para classificação de itens linearmente separáveis, padrões que se encontram em lados opostos por um hiperplano (HAYKIN, 2007).

A Figura 6 apresenta a estrutura de uma RNA do tipo *Perceptron* com n entradas e uma única saída.

Figura 6 – Rede Neural Artificial *Perceptron*.



Fonte: Retirado de (SILVA; SPATTI; FLAUZINO, 2016).

As entradas (x) são ponderadas pelos seus respectivos pesos sinápticos (w) e

combinados (somatório), tendo como valor agregado o bias (θ) necessário para ativação do neurônio. Por último temos a função de ativação, $g(.)$, delimitando o intervalo de saída (y) do neurônio (SILVA; SPATTI; FLAUZINO, 2016).

O funcionamento do neurônio ocorre matematicamente por meio da aplicação das Equações 2.1 e 2.2:

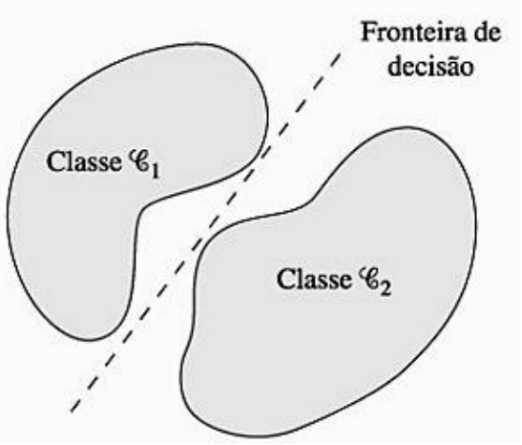
$$u_k = \sum_{i=1}^{m} w_{kj} x_j \tag{2.1}$$

$$y_k = \varphi\left(u_k + b_k\right) \tag{2.2}$$

Onde x_1,x_2,... ,x_m são os sinais de entrada, w_{k1},w_{k2},... ,w_{km} são os pesos sinápticos do neurônio k e y_k é a saída do neurônio. A saída do neurônio é definida através da função de ativação $g(.)$ (SILVA; SPATTI; FLAUZINO, 2016).

O *Perceptron* utiliza o treinamento supervisionado para ajuste dos pesos e limiar, ou seja, para cada amostra dos sinais de entrada se tem a respectiva saída desejada (HAYKIN, 2007).

Figura 7 – Fronteira de decisão de um problema linearmente separável.



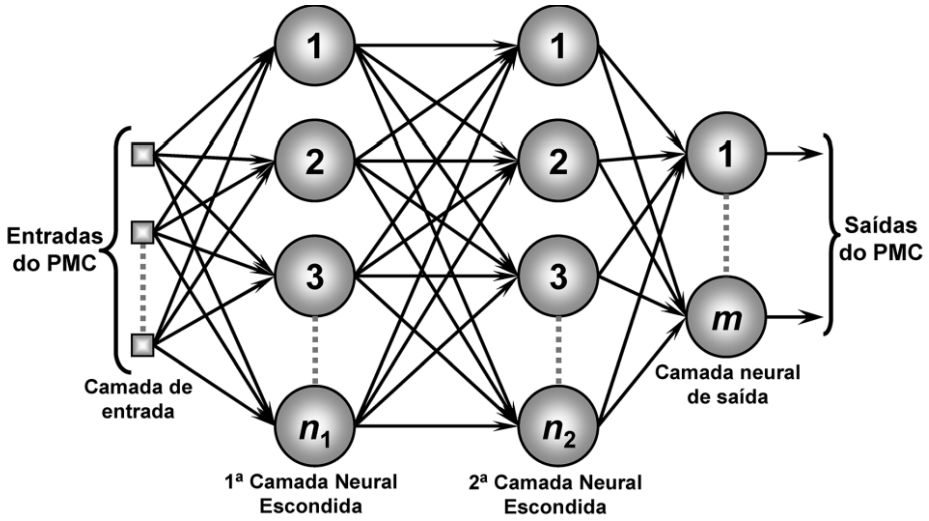
Fonte: Retirado de (HAYKIN, 2007).

2.4.1.3 *Perceptron* de Múltiplas Camadas

Enquanto o *Perceptron* simples só tem uma camada neural, o *Multilayer Perceptron* (MLP) tem, pelo menos, uma camada oculta entre as camadas de entrada e saída da rede (SILVA; SPATTI; FLAUZINO, 2016). Na figura 8 podemos observar as camadas ocultas que permite o aprendizado de tarefas mais complexas, como, por

exemplo, o problema do OU Exclusivo segundo Haykin (2007), tratando classes que não são linearmente separáveis.

Figura 8 – Rede Neural Artificial MLP.



Fonte: Retirado de (SILVA; SPATTI; FLAUZINO, 2016).

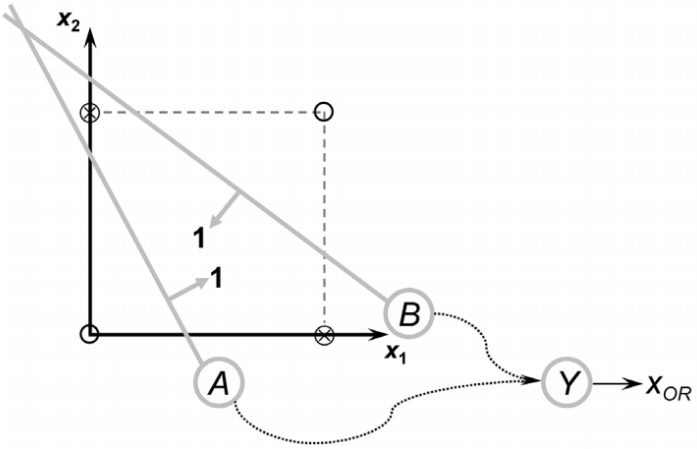
A arquitetura da MLP é organizada em três ou mais camadas: camada de entrada, camadas ocultas e camada de saída. A Camada de entrada é responsável por propagar as entradas da rede para todos os neurônios da camada seguinte. As Camadas ocultas estão localizadas entre a camada de entrada e saída, onde cada neurônio desta camada se comporta como um *Perceptron* simples com saída conectada na entrada dos neurônios da camada seguinte. E a Camada de saída é onde os neurônios processadores estão presentes para classificação (SILVA; SPATTI; FLAUZINO, 2016).

Entre as camadas de entrada e saída, uma ou mais camadas ocultas são normalmente conectadas por matrizes de peso, bias e várias funções de ativação. Essas redes neurais possuem um poder computacional aplicadas em diversos tipos de problemas em diferentes áreas do conhecimento, sendo consideradas uma das arquiteturas mais versáteis quanto à aplicabilidade (HAYKIN, 2007).

Na MLP, os nós são conectados diretamente na camada seguinte, de forma que os sinais seguem um fluxo da entrada até a saída, conhecidas como redes do tipo *feedfoward* (AZEVEDO, 2016). Neste caso o fluxo de informações inicia na camada de entrada, passa pelas camadas intermediárias e finaliza na camada neural de saída e não existe realimentação de informações entre as camadas na rede PMC convencional (SILVA; SPATTI; FLAUZINO, 2016).

Na figura 9, observa-se que apenas com uma camada oculta e dois neurônios nela podemos resolver um problema que não tem classes linearmente separáveis.

Figura 9 – PMC aplicada ao problema XOR.



Fonte: Retirado de (SILVA; SPATTI; FLAUZINO, 2016).

2.4.1.4 Aprendizado das Redes Neurais Artificiais

O treinamento (fase de aprendizado) da RNA, é a etapa onde ocorrem os ajustes dos pesos sinápticos para melhorar o desempenho da rede (HAYKIN, 2007). O aprendizado pode ocorrer de duas formas, supervisionada ou não supervisionada.

O treinamento não supervisionado não possui saída desejada, normalmente, apresenta apenas um único neurônio na camada de saída que envia uma resposta. No aprendizado supervisionado já são conhecidas as saídas desejadas e o treinamento é realizado através da diferença da resposta desejada e a resposta observada na saída da RNA. Essa diferença (erro) vai para o algoritmo de ajuste dos pesos minimizando o erro na próxima iteração (CHAVES, 2019). O algoritmo mais utilizado de propagação de erro é o *backpropagation*.

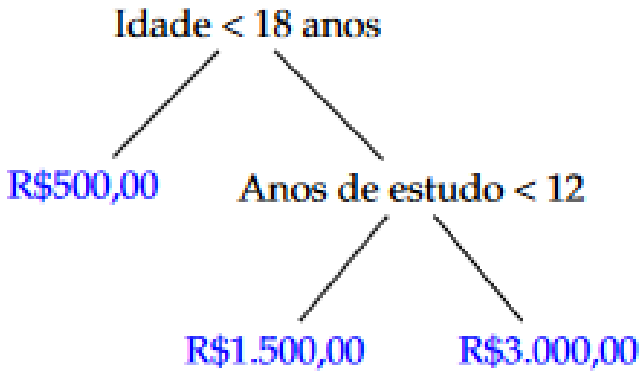
A etapa de *backpropagation*, é o treinamento realizado dentro da rede MLP, consistindo em um método útil no tratamento de informações não linearmente separáveis. Seu funcionamento consiste em duas fases distintas, a propagação (*foward*) e a retro propagação (*backpropagation*). Diferente da fase *foward* na fase do *backpropagation* acontece a atualização dos pesos sinápticos dos neurônios (SILVA; SPATTI; FLAUZINO, 2016). Então, o erro atual é propagado de volta para uma camada anterior, calculando a função de perda. Assim, os gradientes tendem a se tornar menores à medida que o cálculo é estabelecido a cada neurônio, retrocedendo na rede. Dessa forma, os neurônios nas camadas anteriores aprendem de forma muito lenta e o processo de treinamento leva muito mais tempo enquanto a precisão do modelo diminui. Sendo um problema crítico desse tipo de algoritmo (LE et al., 2019). O tempo de treinamento depende da taxa de atualização dos pesos. Se a taxa de atualização dos pesos for muito baixa o algoritmo demora para convergir; no entanto, se ela for alta converge rápido, podendo ficar instável com outras entradas (HAYKIN, 2007).

2.4.2 Decision Trees

Decision Trees é um método de aprendizado de máquina que cria um mapa dos possíveis resultados com base em diferentes escolhas, relacionadas às variáveis do banco de dados (SILVA; VELASQUEZ, 2020). Essas escolhas são organizadas em nós, sendo o primeiro chamado de raiz e os últimos, que indicam os resultados, chamados de folhas (SILVA; VELASQUEZ, 2020). Para prever algo novo, a árvore verifica condições nos nós, seguindo para a esquerda se a condição da raiz é atendida e para a direita, caso contrário. Isso se repete até chegar a uma folha (IZBICKI; SANTOS, 2020).

Na Figura 10, há um exemplo de árvore de decisão que prevê o salário com base na idade e anos de estudo. As árvores são geralmente construídas de cima para baixo, avaliando cada atributo para determinar sua capacidade de classificar os dados (MITCHELL, 1997). Para cada tipo de valor de um atributo, um ramo é criado, e os exemplos de treinamento são classificados nesse ramo. Isso se repete para definir qual atributo é adequado em cada ponto da árvore, cobrindo todos os resultados possíveis (MITCHELL, 1997).

Figura 10 – Exemplo de Decision Trees.



Fonte: Retirado de (IZBICKI; SANTOS, 2020)

A Impureza de Gini, uma medida de desordem nos dados, é usada para escolher o melhor atributo em cada nó (MITCHELL, 1997):

$$H(Q_m) = \sum_k p_{mk}(1 - p_{mk}) \tag{2.3}$$

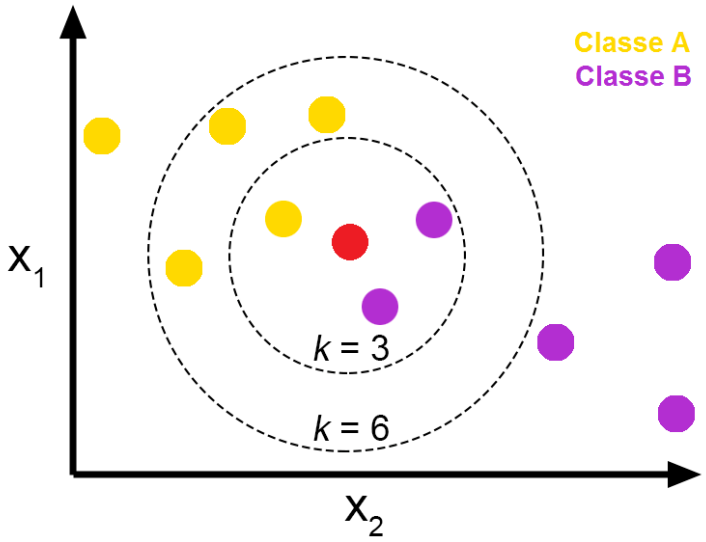
Onde, $H(Q_m)$ é o índice de impureza de Gini, e p_{mk} é a fração de itens do nó m classificado com a classe k (PEDREGOSA et al., 2011). Essa métrica ajuda na classificação de observações usando múltiplos atributos.

2.4.3 K-Nearest Neighbors

O método dos k-vizinhos mais próximos (KNN) é uma técnica de aprendizado de máquina baseada em instância, enquadrando-se na categoria de aprendizado não generalizante. Em vez de construir um modelo interno geral, o KNN armazena instâncias dos dados de treinamento, calculando a classificação por meio de um voto de maioria dos vizinhos mais próximos de cada ponto (PEDREGOSA et al., 2011). Essa abordagem difere das discutidas anteriormente, permitindo a construção de uma aproximação diferente para a função de destino para cada instância de consulta distinta que deve ser classificada (MITCHELL, 1997).

O KNN é um dos métodos mais conhecidos dessa abordagem. Esse algoritmo realiza a classificação com base na distância (geralmente euclidiana) de uma nova observação em relação às demais, considerando um parâmetro k que representa o número de vizinhos selecionados (IZBICKI; SANTOS, 2020; SILVA; VELASQUEZ, 2020). Se, por exemplo, k for igual a 3, a instância a ser classificada será comparada às três mais próximas.

Figura 11 – Exemplo do algoritmo KNN.



Fonte: Retirado de (IZBICKI; SANTOS, 2020)

Altamente eficaz para sistemas práticos, o KNN demonstra robustez diante de dados de treinamento ruidosos e eficiência quando fornecido um conjunto suficientemente grande de dados de treinamento (MITCHELL, 1997). No entanto, é importante destacar que o k-NN utiliza todos os atributos de uma instância para calcular a distância, podendo levar a resultados enganosos quando muitos desses atributos não são relevantes para a classificação (MITCHELL, 1997). Para contornar esse problema, podem ser atribuídos pesos diferentes a cada atributo ou até mesmo eliminar aqueles considerados menos relevantes (MITCHELL, 1997).

O KNN funciona por meio da memorização do conjunto de treinamento. Quando um novo dado é inserido, ele recebe o rótulo de seus vizinhos mais próximos. A proximidade é definida pela distância euclidiana dos vetores de atributos(Eq 2.4), onde x_i e x_j representam objetos em um determinado espaço e x_i^l e x_j^l são elementos desses vetores correspondentes aos valores da coordenada l (atributos) (SILVA; VELASQUEZ, 2020). O valor de k, um número ímpar maior que 1, é crucial para determinar o rótulo atribuído ao novo dado, sendo baseado na ocorrência entre os k vetores de atributos com a menor distância em relação ao novo dado.

$$d\left(\mathbf{x}_i, \mathbf{x}_j\right) = \sqrt{\sum_{l=1}^{d} \left(x_i^l - x_j^l\right)^2} \tag{2.4}$$

2.4.4 AdaBoost

O AdaBoost em *machine learning* é uma técnica de modelagem preditiva que pertence à categoria de métodos de *Ensemble*. Essa abordagem, também conhecida como Adaptive Boosting, utiliza principalmente árvores de decisão com apenas um nível, conhecidas como *Decision Stumps*. O AdaBoost inicia a construção de um modelo, atribuindo pesos iguais a todos os pontos de dados, e, em seguida, aplica pesos maiores aos pontos classificados incorretamente. Esse processo é repetido em modelos subsequentes até que um erro menor seja alcançado.

Em termos mais simples, o AdaBoost é uma técnica popular de Boosting, que combina vários modelos de aprendizado fracos sequencialmente, cada um tentando corrigir as falhas do seu antecessor. Ao prestar atenção às instâncias onde o modelo anterior falhou, o AdaBoost constrói novos preditores capazes de corrigir esses erros. Qualquer algoritmo de machine learning pode ser usado como o modelo de aprendizagem fraco inicial, contanto que aceite pesos no conjunto de treinamento.

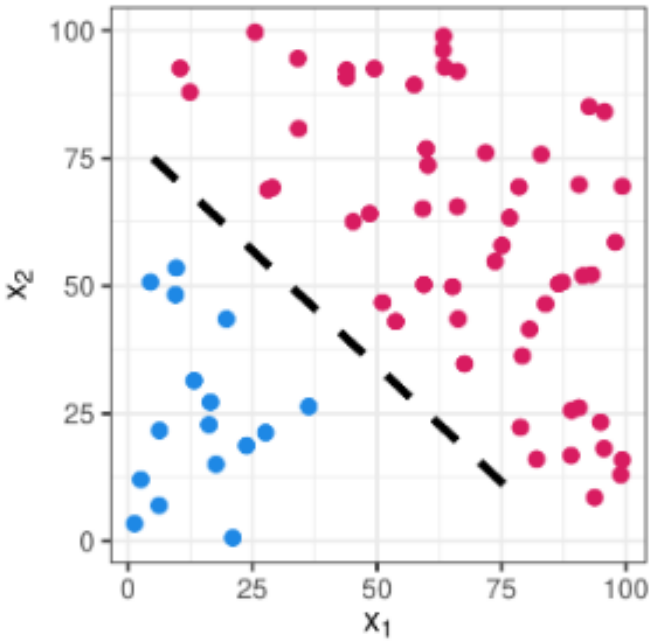
2.4.5 Support Vector Machine

Support Vector Machine (SVM) representam uma abordagem proposta por Cortes e Vapnik (1995) que consiste em mapear os dados de entrada para um espaço de alta dimensão por meio de um mapeamento não linear predefinido. Nesse espaço, é construída uma superfície de decisão linear chamada hiperplano, com propriedades especiais que garantem uma alta capacidade de generalização da rede, permitindo assim a classificação eficiente dos dados (CORTES; VAPNIK, 1995).

A Figura 12 exemplifica um conjunto de dados separado por um hiperplano de forma linear. Quando há vários hiperplanos que separam o espaço perfeitamente, a SVM busca aquele que possui a maior margem, ou seja, o que fica mais distante de todos

os pontos observados (IZBICKI; SANTOS, 2020). Os pontos utilizados para definir as margens são denominados vetores de suporte.

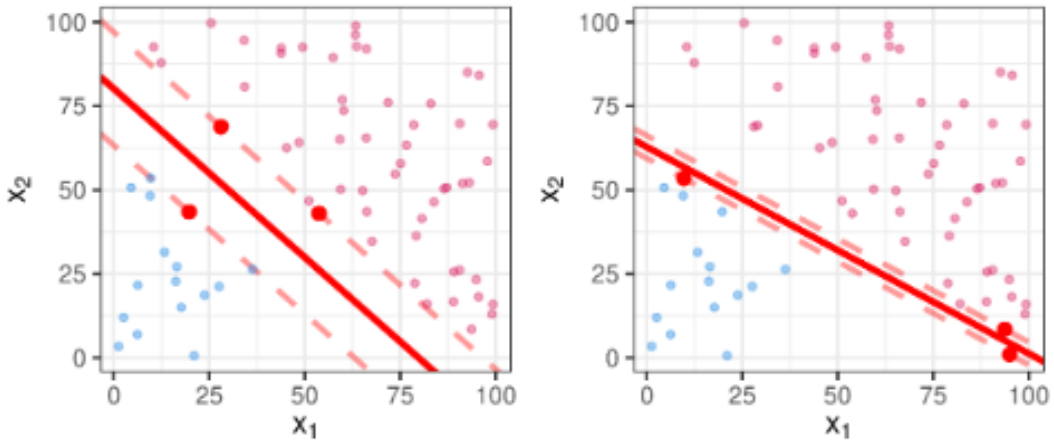
Figura 12 – Conjunto de dados perfeitamente separados por um hiperplano.



Fonte: Retirado de (IZBICKI; SANTOS, 2020)

É importante observar que esse algoritmo é bastante sensível a pequenas mudanças nos dados, como ilustrado na Figura 13.

Figura 13 – Alteração nos vetores de suporte e margem devido a somente uma observação.



Fonte: Retirado de (IZBICKI; SANTOS, 2020)

As SVMs lineares são eficazes na classificação de conjuntos de dados que são linearmente separáveis ou possuem uma distribuição aproximadamente linear. No entanto, há muitos casos em que não é possível dividir satisfatoriamente os dados de treinamento

por um hiperplano. Para superar essa limitação, é feito um mapeamento dos dados para um espaço de maior dimensão, tornando-os linearmente separáveis nesse espaço (LORENA; CARVALHO, 2007). Como o cálculo desse mapeamento pode ser custoso ou inviável, são aplicadas funções Kernel, amplamente utilizadas devido à simplicidade de seu cálculo e à capacidade de representar espaços abstratos (LORENA; CARVALHO, 2007).

Um Kernel K é uma função que recebe dois pontos x_i e x_j do espaço de entrada e calcula o produto escalar desses dados no espaço de características (HERBRICH, 2001). Entre os Kernels mais utilizados estão os Polinomiais, Gaussianos ou RBF (Radial-Basis Function) e Sigmoidais, representados pelas Equações 2.5, 2.6 e 2.7 (Lorena, 2007).

$$K(x_i, x_j) = (\delta(x_i \cdot x_j) + k)^d \tag{2.5}$$

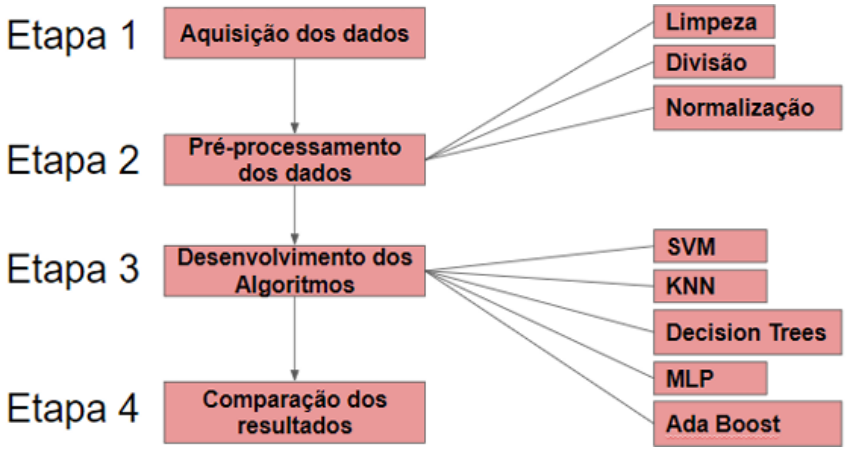
$$K(x_i, x_j) = \exp(-\sigma||x_i - x_j||^2) \tag{2.6}$$

$$K(x_i, x_j) = \tanh(\delta(x_i \cdot x_j) + k) \tag{2.7}$$

3 MATERIAIS E MÉTODOS

O presente estudo apresenta-se como uma pesquisa exploratória com características predominantemente qualitativas. Este trabalho tem como finalidade a realização de um estudo visando classificar os sedimentos marinhos da região da Margem Equatorial Brasileira, mediante dados sísmicos, dados sedimentológicos e algoritmos de aprendizado de máquina. Neste trabalho, são comparados cinco algoritmos de aprendizagem de máquina na classificação de sedimentos do fundo do mar na Margem Equatorial Brasileira. O estudo foi dividido em quatro etapas, conforme ilustrado na Figura 14. A seguir, são apresentados os seguintes tópicos: aquisição da base de dados, pré-processamento dos dados, desenvolvimento dos algoritmos e a comparação dos resultados encontrados.

Figura 14 – Etapas da Metodologia.



Fonte: Autor.

3.1 Base de dados

Na primeira etapa do trabalho, foram reunidos 5.515 dados de sedimentos superficiais provenientes do Banco Nacional de Dados Oceanográficos (BNDO) e do Banco de Dados da Indústria do Petróleo (BAMPETRO). Esses dados apresentam duas informações básicas: a natureza do fundo oceânico, determinada por descrição visual, e dados sedimentológicos processados em laboratório. Os teores de tamanho de grão foram classificados conforme a classificação ternária de Shepard em três classes principais: Lama, Areia e Cascalho. Além disso, o teor de carbonato de cálcio foi categorizado em quatro classes: Litoclástico ($CaCO3 < 30\%$), Litobioclástico ($30\% < CaCO3 < 50\%$), Litobioclástico ($50\% < CaCO3 < 70\%$) e Litoclástico ($CaCO3 > 70\%$) (LARSONNEUR, 1977).

Os dados de atributos sísmicos consistem em resultados da interpretação e processamento de 384 linhas públicas de sísmica 2D multicanal, extraídas no software

Paleoscan. Os primeiros 500 milissegundos de cada linha sísmica foram interpretados manualmente para a extração de cinco atributos sísmicos: amplitude, intensidade de reflexão, fase, frequência e coerência.

O resultado desta etapa foi um arquivo ".csv (Comma-Separated Values)"contendo 5.515 linhas referentes aos pontos de sedimento e 106 colunas abrangendo diversos parâmetros sedimentológicos.

Na etapa seguinte, realizou-se o pré-processamento desses dados para prepará-los como entrada para os algoritmos de aprendizagem de máquina.

3.2 Pré-processamento dos dados

Após a coleta inicial, o arquivo da etapa anterior continha 5.512 linhas e 106 colunas. Inicialmente, na etapa 2 do trabalho, foram separados apenas os dados de interesse. Isso incluiu as colunas com informações sobre a Classe Shepard (Areia, Cascalho, Lama), a latitude e a longitude do sedimento, bem como os cinco atributos sísmicos associados.

Na Figura 15, as células selecionadas exemplificam sedimentos que foram classificados, mas não possuíam uma linha sísmica associada. Nesses casos, as linhas da base foram excluídas. O resultado desse processo foi um arquivo ".csv"com 8 colunas, contendo 101 linhas de areia, 74 linhas de cascalho e 73 linhas de lama. Esse pré-processamento foi essencial para garantir que apenas os dados relevantes e associados à informação sísmica fossem considerados nas etapas subsequentes da análise.

Figura 15 – Base de dados coletados.

NM_TENCA_A	X_UTM	Y_UTM	AMPLITUDE_	RAS_INSTFR	RAS_INSTPH	RAS_COHERE	RAS_INTENS
AREIA	1595434.5424	9440577.1051	719.7239990	0.0192309	0.1688510	0.9131300	23.3439007
AREIA	1591625.7695	9436707.8771	-9999.0000000	-9999.0000000	-9999.0000000	-9999.0000000	-9999.0000000
AREIA	1591620.1144	9436334.0449	-9999.0000000	-9999.0000000	-9999.0000000	-9999.0000000	-9999.0000000
AREIA	1567542.9925	9456314.2389	82.1262970	0.0160263	0.3073120	0.8710590	8.5595303
AREIA	1507061.5831	9457153.0757	936.6740112	0.0284669	-0.0155058	0.8873850	25.6208992
AREIA	1507049.0181	9456220.4552	947.3629761	0.0284511	-0.0130500	0.8899530	25.8560009
AREIA	275995.4162	10070045.8591	276.6029968	0.0200214	0.1815200	0.8869300	15.1597004
AREIA	1238157.0482	9589101.2481	457.3190002	0.0240550	0.2201370	0.9830650	22.4715996
	1099088.1009	9690955.3354	270.7980042	0.0181933	0.1301570	0.8721130	15.6310997
AREIA	830849.2208	9752536.7835	239.3609924	0.0211553	-0.0072241	0.8513240	13.3385000
	597976.0634	9901984.9258	-449.8349915	0.0217886	-0.8941990	0.8389660	20.4855995
AREIA	613496.4904	9915615.2582	73.9372025	0.0170331	0.1266800	0.8492690	8.8279305

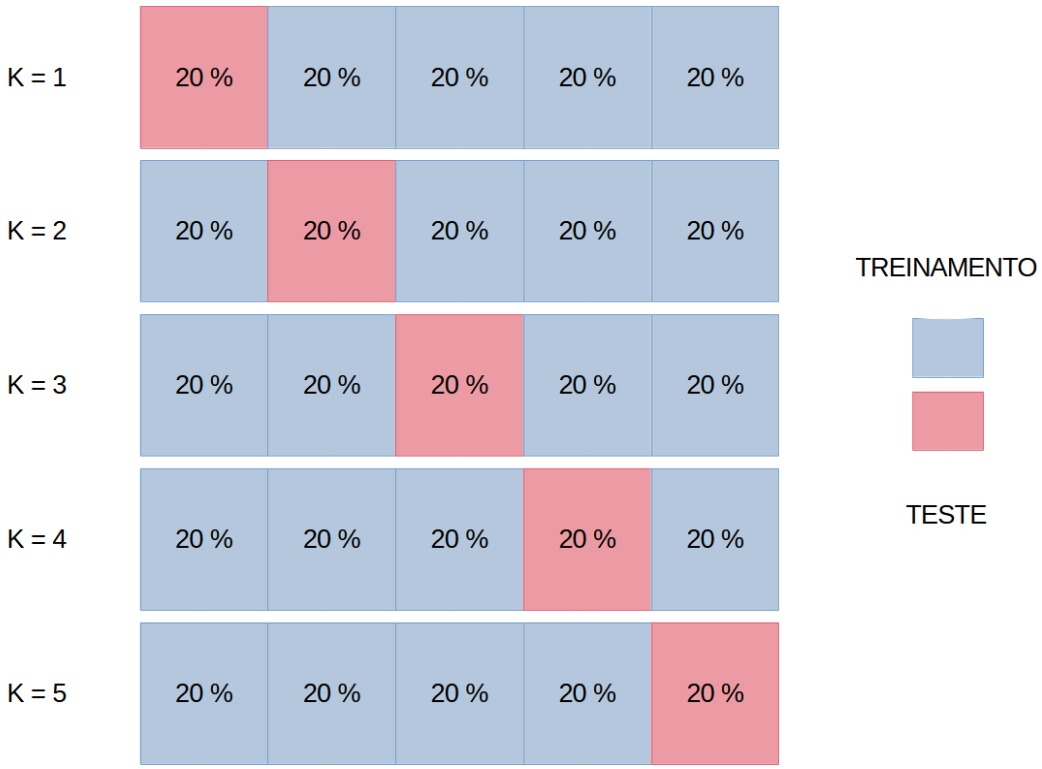
Fonte: Autor.

A Figura 16 abaixo ilustra o procedimento de Validação Cruzada dos dados, uma técnica amplamente utilizada em aprendizado de máquina para avaliar o desempenho dos algoritmos.

A Validação Cruzada envolve a divisão dos dados em subconjuntos e a alteração desses subconjuntos à medida que o algoritmo é executado. Se K for igual a 5, isso significa que os dados são divididos em 5 partes (20% dos dados cada), e o algoritmo é executado

cinco vezes. Na primeira execução, a Parte 1 é designada para teste e o restante para treinamento. Na segunda execução, a Parte 2 dos dados é destinada para teste, e assim por diante até K = 5.

Figura 16 – Divisão dos dados para a Validação Cruzada.



Fonte: Autor.

Esse método ajuda a garantir uma avaliação mais robusta do modelo, uma vez que ele é testado em diferentes subconjuntos de dados, reduzindo a dependência de uma única divisão específica dos dados. Isso proporciona uma avaliação mais geral do desempenho do algoritmo.

A técnica de Validação Cruzada é utilizada para avaliar a capacidade de generalização de um modelo, com ela é possível realizar o teste do algoritmo em todos os dados. No final das cinco execuções é realizado a média dos resultados encontrados e obtido o desempenho final do algoritmo.

3.3 Aplicação dos algoritmos

Nesta fase, foram aplicados cinco algoritmos de aprendizado de máquina para classificação de sedimentos marinhos. Cada um desses algoritmos desempenha um papel crucial na análise e categorização eficiente dos dados sedimentares.

O Multilayer Perceptron (MLP) é uma rede neural artificial que compreende uma ou mais camadas ocultas, juntamente com um número variável de neurônios. Derivado

do algoritmo Perceptron, o MLP permite uma flexibilidade considerável na configuração, com o número de camadas ocultas e neurônios ajustáveis de acordo com a complexidade do problema. A sensibilidade ao problema aumenta com o acréscimo de neurônios, embora seja importante considerar o risco de overfitting (MURTAGH, 1991).

As Decision Trees, ou Árvores de Decisão, são estruturas de árvores que facilitam a tomada de decisões com base em condições e ramificações. Cada nó na árvore representa uma decisão baseada em um atributo específico, culminando em folhas que indicam as classes ou resultados finais. Esta abordagem visual e intuitiva é eficaz para problemas de classificação.

O K-Nearest Neighbors (KNN) é um método que classifica uma instância com base nas classes das instâncias vizinhas, sendo a proximidade determinada por uma métrica de distância, geralmente euclidiana. Este algoritmo é eficaz em identificar padrões em dados próximos no espaço de características.

O AdaBoost é uma técnica de boosting que combina múltiplos modelos fracos sequencialmente para formar um modelo forte. Cada modelo subsequente é treinado para corrigir as deficiências do modelo anterior, melhorando assim a capacidade de predição global.

Finalmente, o *Support Vector Machine* (SVM) é aplicado para problemas de classificação e regressão, buscando identificar um hiperplano ótimo que separa eficientemente os dados em diferentes classes. Essa abordagem é especialmente útil quando se trata de conjuntos de dados complexos e não linearmente separáveis.

O algoritmo *Suport Vector Machine* é aplicado geralmente em problemas de classificação e regressão, tem como ideia principal encontrar um hiperplano que melhor separa os dados em classes diferentes.

3.4 Comparação dos resultados

Uma matriz de confusão é uma tabela que indica os erros e acertos do modelo. Eles descrevem o resultado da classificação de cada registro, e através deles podemos encontrar as demais métricas (DIAS; PASCUTTI; SILVA, 2016).

Para avaliação dos resultados as seguintes métricas de desempenho foram utilizadas: *Accuracy*, *Precision*, *Recall* e *F1-Score*. No entanto é preciso saber os seguintes valores antes:

- *True Positive* (TP) é a classificação correta da classe Positivo.
- *True Negative* (TN) é a classificação correta da classe Negativo.

- *False Positive* (FP) é o erro em que o modelo previu a classe Positivo quando o valor real era classe Negativo.
- *False Negative* (FN) é o erro em que o modelo previu a classe Negativo quando o valor real era classe Positivo.

3.4.1 Accuracy

A *accuracy* é a taxa de acerto, representando as instâncias corretamente classificadas dentre todas as classificadas, podendo ser calculada através da seguinte equação:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{3.1}$$

Em trabalhos sobre aprendizado de máquinas, a acurácia é uma métrica amplamente utilizada, pois revela o percentual de acertos do modelo. No entanto, é importante destacar que, embora seja uma métrica relevante, a acurácia por si só pode não fornecer uma visão completa do desempenho geral do modelo (COSTA, 2018).

A acurácia é calculada como a razão entre o número total de predições corretas (verdadeiros positivos e verdadeiros negativos) e o número total de predições. No entanto, em situações desbalanceadas, onde uma classe é dominante em relação às outras, a acurácia pode ser enganadora. Por exemplo, se um modelo classificar todas as instâncias como pertencentes à classe dominante, ainda assim terá uma alta acurácia, mas seu desempenho real em termos de identificar outras classes pode ser inadequado.

Portanto, é fundamental complementar a análise com métricas adicionais, como precisão, recall e F1-Score, para obter uma compreensão mais abrangente do desempenho do modelo, especialmente em contextos desbalanceados.

3.4.2 Precision

A *precision* traduz o quanto o modelo é capaz de detectar as amostras positivas, podendo ser calculada através da seguinte equação:

$$Precision = \frac{TP}{TP + FP} \tag{3.2}$$

Dentre todas as classificações de classe Positivo que o modelo fez, quantas estão corretas. É a porcentagem de acerto que realmente são acertos (COSTA, 2018).

Essa métrica é valiosa para avaliar a confiabilidade das predições positivas feitas pelo modelo. Uma precisão alta indica que, quando o modelo classifica uma instância como positiva, ela tem uma alta probabilidade de estar correta. No entanto, a precisão, por si só, pode não ser suficiente para avaliar completamente o desempenho do modelo, especialmente em casos desbalanceados, onde a classe positiva é menos representada. Portanto, é comum analisar a precisão em conjunto com outras métricas, como recall e F1-Score.

3.4.3 Recall

Ao contrário da *precision* a *recall* é capaz de discernir as amostras que são negativas, obtida através dos casos que os sedimentos não são classificados corretamente (BORCHARTT, 2013), podendo ser calculada através da seguinte equação:

$$Recall = \frac{TP}{TP + FN} \tag{3.3}$$

A recall representa a porcentagem de instâncias da classe Positivo que foram corretamente identificadas pelo modelo em relação ao total de instâncias que realmente são positivas. Em outras palavras, ela expressa a capacidade do modelo de recuperar todas as instâncias relevantes da classe Positivo. Assim como a precisão, a recall é uma métrica importante, especialmente em situações em que a identificação correta dos casos positivos é crucial. No entanto, em alguns contextos, pode ser necessário equilibrar a precisão e a recall, o que é abordado pela métrica F1-Score.

3.4.4 *F1-Score*

É uma maneira de observar somente 1 métrica ao invés de duas (*precision* e *recall*) em alguma situação. É uma média harmônica entre as duas, ou seja, quando tem-se um *F1-Score* baixo, é um indicativo de que *precision* ou *recall* está baixo.

$$F1Score = \frac{2 * Precision * Recall}{Precision + Recall} \tag{3.4}$$

O F1-Score é particularmente útil em situações em que há um desequilíbrio entre as classes, ou seja, quando uma classe é muito mais numerosa do que a outra. Ele

fornece uma métrica única que combina tanto a precisão quanto a recall em uma única medida, sendo uma média harmônica que penaliza valores extremos. Portanto, um F1-Score mais alto indica um bom equilíbrio entre precisão e recall.

3.5 Ambiente de teste

Para realização do trabalho foi utilizado a linguagem de programação python com as bibliotecas pandas, csv e sklearn, no ambiente do Google Colab Pro, que disponibiliza um hardware contendo GPU Tesla P100-PCIE-16GB, CPU Intel Xeon @ 2x 2GH, Memória RAM 13GB, HD 70GB, Sistema Operacional Ubuntu 18.04 Linux 4.19.104.

O Google Colab Pro é um ambiente de notebooks Jupyter que não requer configuração e é executado na nuvem. Não é preciso instalar o Python e já vem com uma enorme variedade de bibliotecas instaladas em um contêiner de notebook Jupyter.

Pode-se escrever e executar códigos em Python diretamente no navegador. É muito indicado para Aprendizado de Máquina, Inteligência Artificial e Ciência de Dados, entre outras áreas de conhecimento.

4 RESULTADOS E DISCUSSÃO

Os testes realizados abrangeram todos os classificadores mencionados no capítulo anterior. Os algoritmos de aprendizado de máquina possuem vários parâmetros que podem influenciar no aprendizado (treinamento), resultando em desempenhos satisfatórios ou inaceitáveis. Por isso, é importante encontrar a melhor combinação desses parâmetros para garantir a confiabilidade na classificação dos sedimentos.

Dessa forma, foram realizados testes com os dados de latitude, longitude do sedimento e os cinco atributos sísmicos: amplitude, intensidade de reflexão, fase, frequência e coerência. Mas os melhores resultados foram encontrados utilizando latitude, longitude e apenas três dos atributos sísmicos: amplitude, fase e frequência. Todos os resultados são apresentados abaixo, na seguinte ordem de classificadores: MLP, *Decision Trees*, KNN, *AdaBoost* e SVM.

A análise dos resultados da classificação de sedimentos marinhos (Tabela 1) revelou características interessantes sobre o desempenho dos diferentes algoritmos. Foram conduzidos testes utilizando dados de latitude, longitude e cinco atributos sísmicos, sendo amplitude, intensidade de reflexão, fase, frequência e coerência. No entanto, os resultados mais promissores foram alcançados ao empregar latitude, longitude e apenas três dos atributos sísmicos: amplitude, fase e frequência.

Tabela 1 – Resultados de cada classificador.

Classificador	**Accuracy**	**Precision**	**Recall**	**F1-Score**
Decision Trees	76,63%	79,33%	76,63%	77,92%
MLP	83,99%	84,56%	83,99%	84,27%
SVM	83,21%	84,11%	83,21%	83,66%
KNN	73,22%	78,47%	73,22%	75,74%
AdaBoost	74,38%	72,82%	74,38%	73,35%

Fonte: Autor.

A Tabela 2 abaixo, apresenta os resultados do classificador MLP. O algoritmo foi implementado com as seguintes configurações: 50 camadas ocultas, 100 neurônios por camadas, taxa de aprendizado 0,000001, função de ativação RELU e otimização ADAM. O resultado da rede neural foi em média de 83,99% de acurácia, 84,56% de precisão, 83,99% para recall e 84,27% em F1-score. Os melhores valores de *Accuracy* (91,84%) encontrados no classificador MLP foram obtidos quando K = 3.

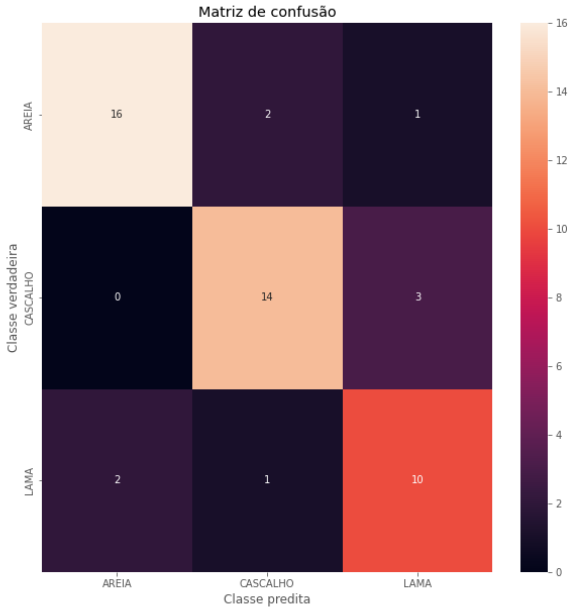
Tabela 2 – Resultados do Classificador MLP.

K-FOLD	Accuracy	Precision	Recall	F1-Score
K = 1	81,63%	81,99%	81,63%	81,81%
K = 2	79,59%	80,90%	79,59%	80,24%
K = 3	91,84%	92,15%	91,84%	91,99%
K = 4	81,19%	81,37%	81,19%	81,28%
K = 5	85,71%	86,38 %	85,71%	86,05%
MÉDIA	**83,99%**	**84,56%**	**83,99%**	**84,27%**

Fonte: Autor.

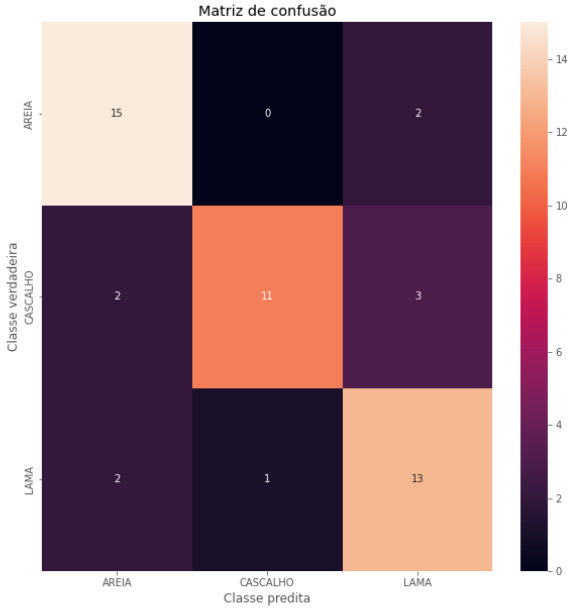
Nas figuras 17, 18, 19, 20 e 21 estão apresentadas as matrizes de confusão de cada K-fold atribuído para validação dos dados.

Figura 17 – Matriz de Confusão MLP quando K-Fold = 1.



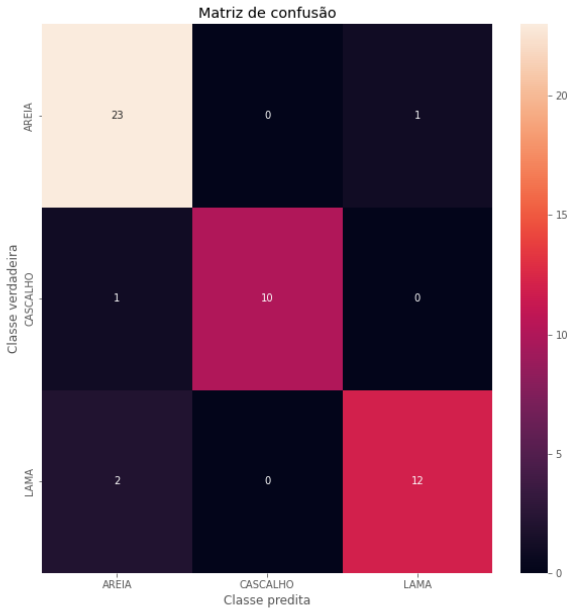
Fonte: Autor.

Figura 18 – Matriz de Confusão MLP quando K-Fold = 2.



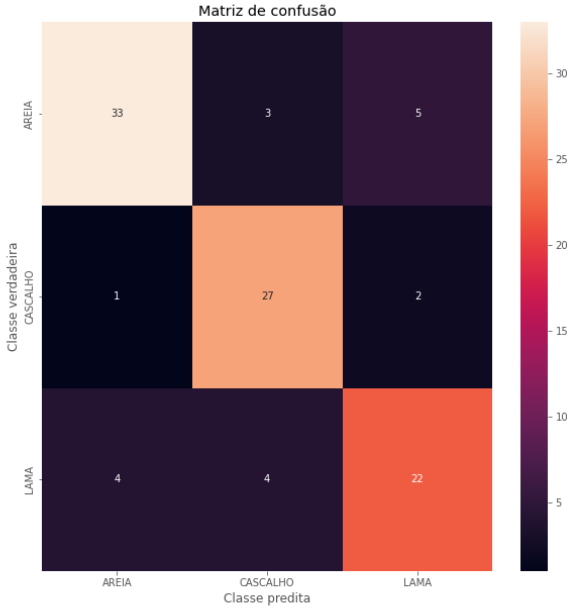
Fonte: Autor.

Figura 19 – Matriz de Confusão MLP quando K-Fold = 3.



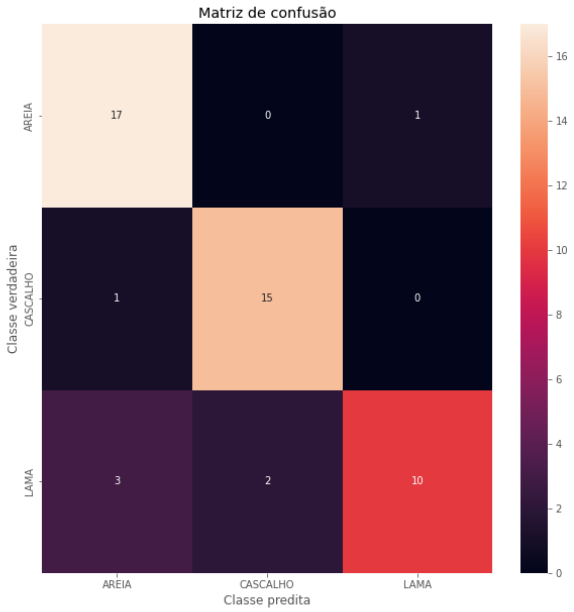
Fonte: Autor.

Figura 20 – Matriz de Confusão MLP quando K-Fold = 4.



Fonte: Autor.

Figura 21 – Matriz de Confusão MLP quando K-Fold = 5.



Fonte: Autor.

Na Tabela 3, são apresentados os resultados do classificador *Decision Trees*.

Sua implementação foi realizada com as configurações padrões, conforme encontrada na biblioteca *Sklearn*. O resultado da árvore de decisão foi em média de 76,63% de acurácia, 79,33% de precisão, 76,63% para recall e 77,92% em F1-score. Os melhores valores de *Accuracy* (83,67%) encontrados no classificador árvore de decisão foram obtidos quando K = 3 e K = 5 com 76,63% de acurácia.

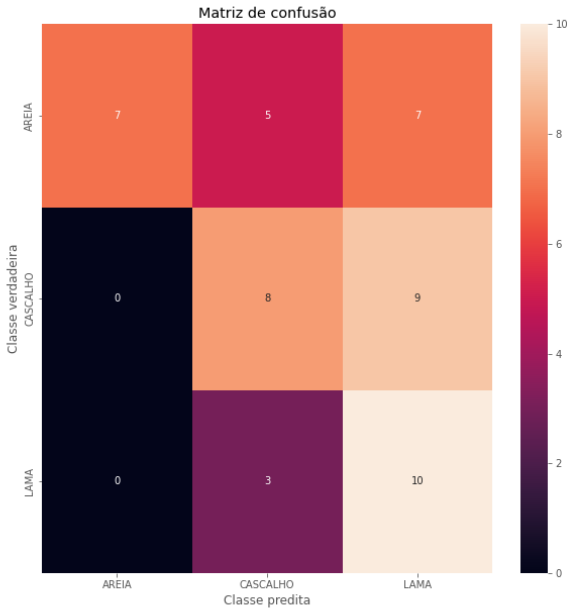
Tabela 3 – Resultados do Classificador *Decision Trees*.

K-FOLD	**Accuracy**	**Precision**	**Recall**	**F1-Score**
K = 1	59,18%	67,19%	59,18%	62,93%
K = 2	73,47%	74,75%	73,47%	74,10%
K = 3	83,67%	85,13%	83,67%	84,40%
K = 4	83,17%	83,02%	83,17%	83,09%
K = 5	83,67%	86,57%	83,67%	85,10%
MÉDIA	**76,63%**	**79,33%**	**76,63%**	**77,92%**

Fonte: Autor.

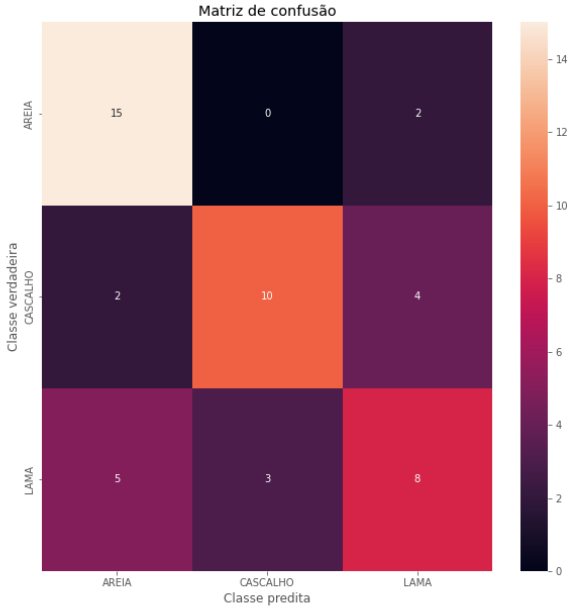
Nas figuras 22, 23, 24, 25 e 26 estão apresentadas as matrizes de confusão de cada K-fold atribuído para validação dos dados.

Figura 22 – Matriz de Confusão Decision Trees quando K-Fold = 1.



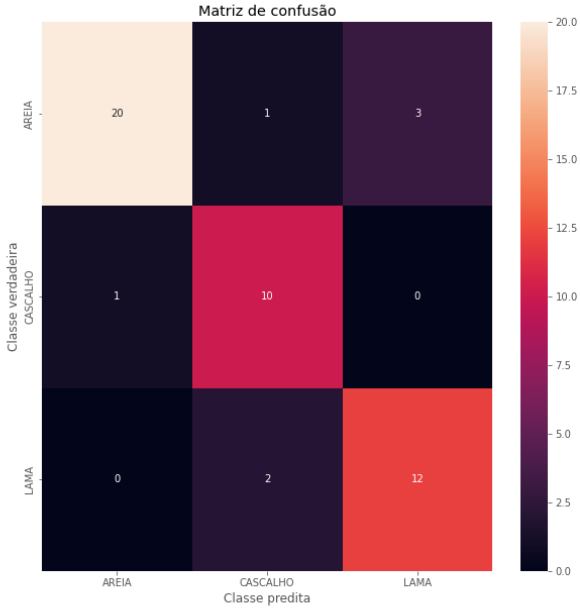
Fonte: Autor.

Figura 23 – Matriz de Confusão Decision Trees quando K-Fold = 2.



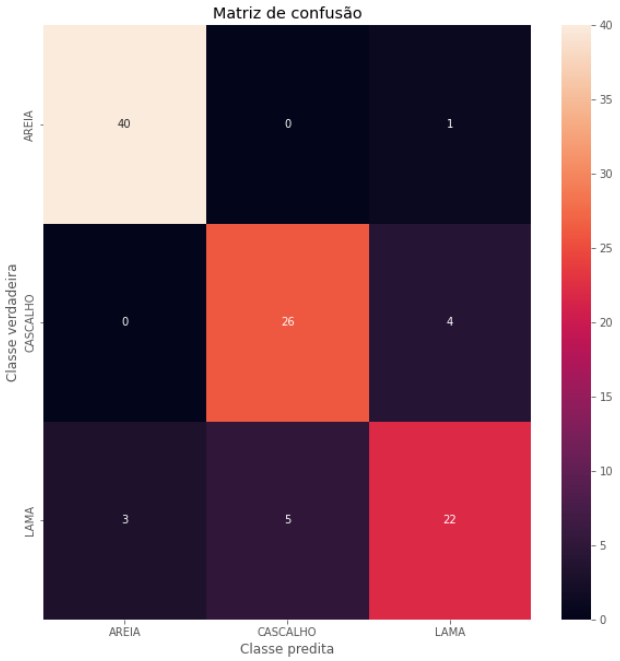
Fonte: Autor.

Figura 24 – Matriz de Confusão Decision Trees quando K-Fold = 3.



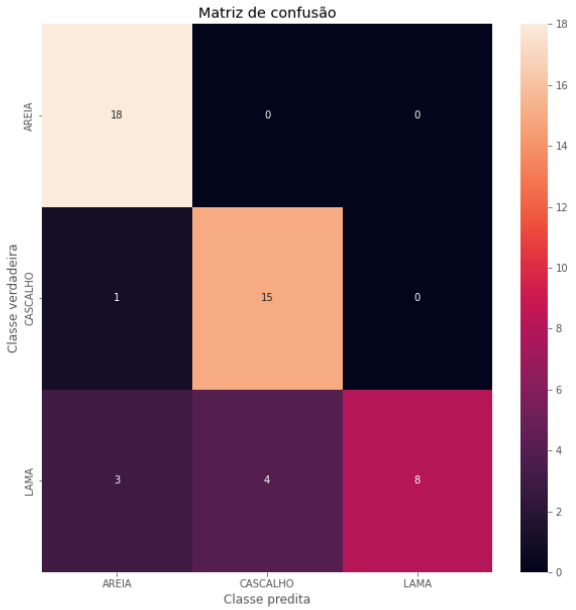
Fonte: Autor.

Figura 25 – Matriz de Confusão Decision Trees quando K-Fold = 4.



Fonte: Autor.

Figura 26 – Matriz de Confusão Decision Trees quando K-Fold = 5.



Fonte: Autor.

A Tabela 4 abaixo, apresenta os resultados do classificador KNN. O algoritmo foi implementado com a seguinte configuração: o valor de K = 3, ou seja, os três vizinhos mais próximos são utilizados para classificação dos sedimentos. O resultado do KNN foi em média de 75,22% de acurácia, 78,47% de precisão, 73,22% para recall e 75,74% em F1-score. Os melhores valores de *Accuracy* (83,67%) encontrados no classificador KNN foram obtidos quando K = 1 com 73,22% de acurácia.

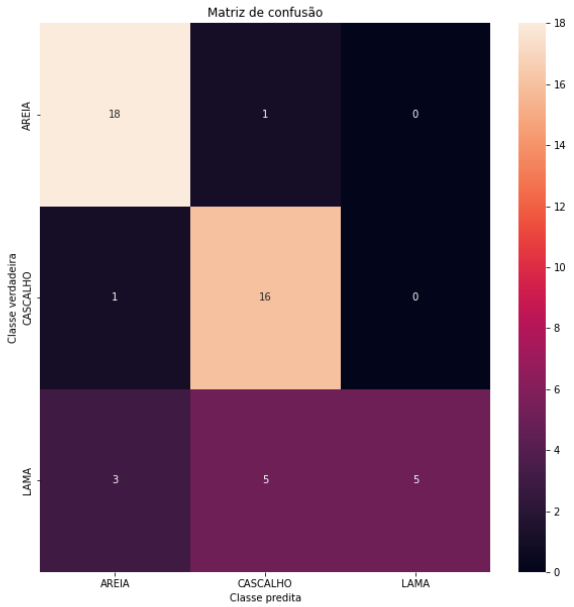
Tabela 4 – Resultados do Classificador KNN.

K-FOLD	**Accuracy**	**Precision**	**Recall**	**F1-Score**
K = 1	79,59%	83,49%	79,59%	81,49%
K = 2	65,31%	72,62%	65,31%	68,77%
K = 3	77,55%	81,85%	77,55%	79,64%
K = 4	74,26%	77,70%	74,26%	75,94%
K = 5	69,39%	76,67%	69,39%	72,85%
MÉDIA	**73,22%**	**78,47%**	**73,22%**	**75,74%**

Fonte: Autor.

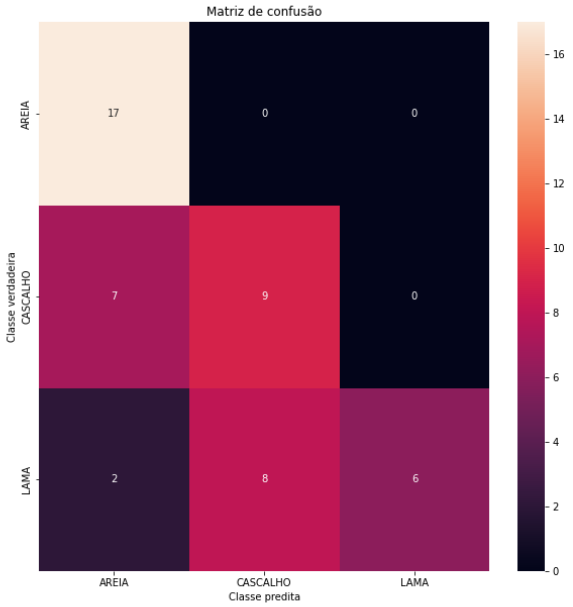
Para fins comparativos, nas figuras 27, 28, 29, 30 e 31 estão apresentadas as matrizes de confusão de cada K-fold atribuído para validação dos dados.

Figura 27 – Matriz de Confusão KNN quando K-Fold = 1.



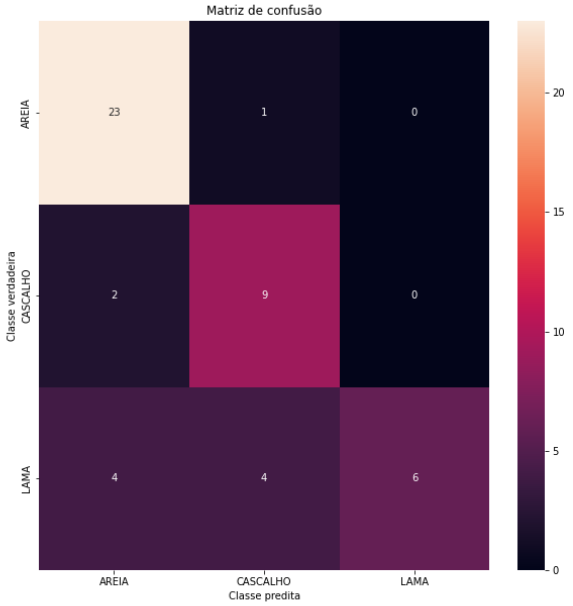
Fonte: Autor.

Figura 28 – Matriz de Confusão KNN quando K-Fold = 2.



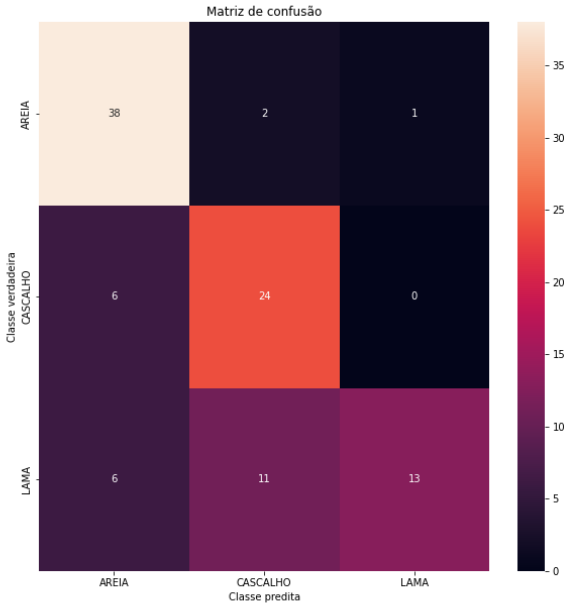
Fonte: Autor.

Figura 29 – Matriz de Confusão KNN quando K-Fold = 3.



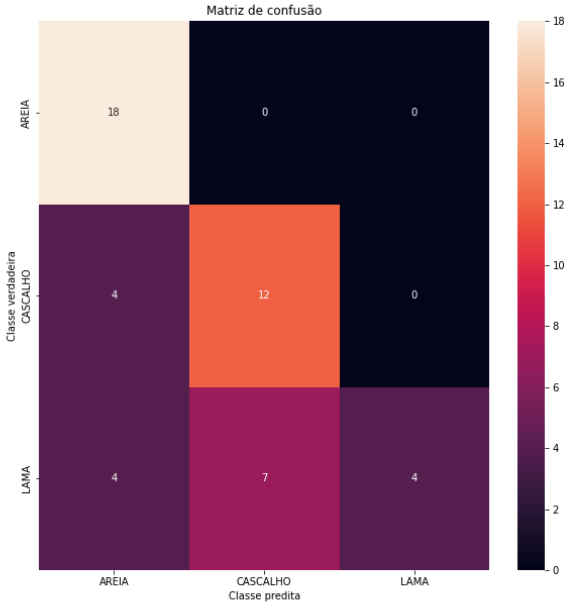
Fonte: Autor.

Figura 30 – Matriz de Confusão KNN quando K-Fold = 4.



Fonte: Autor.

Figura 31 – Matriz de Confusão KNN quando K-Fold = 5.



Fonte: Autor.

A Tabela 5 abaixo, apresenta os resultados do classificador *AdaBoost*. Sua

implementação foi realizada com as configurações padrões, conforme encontrada na biblioteca *Sklearn*. O resultado do AdaBoost foi em média de 73,22% de acurácia, 74,38% de precisão, 74,38% para recall e 73,35% em F1-score. Os melhores valores de *Accuracy* (73,22%) encontrados no classificador AdaBoost foram obtidos quando K = 1 com 73,22% de acurácia.

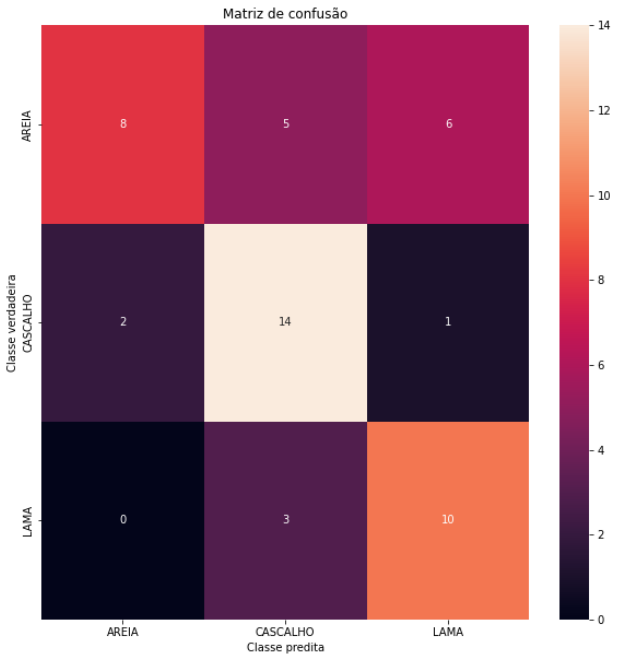
Tabela 5 – Resultados do Classificador *AdaBoost*.

K-FOLD	**Accuracy**	**Precision**	**Recall**	**F1-Score**
K = 1	65,31%	68,70%	65,31%	66,96%
K = 2	57,14%	41,73%	57,14%	48,24%
K = 3	85,71%	88,98%	85,71%	87,32%
K = 4	84,16%	85,11%	84,16%	84,63%
K = 5	79,59%	79,59%	79,59%	79,59%
MÉDIA	**74,38%**	**72,82%**	**74,38%**	**73,35%**

Fonte: Autor.

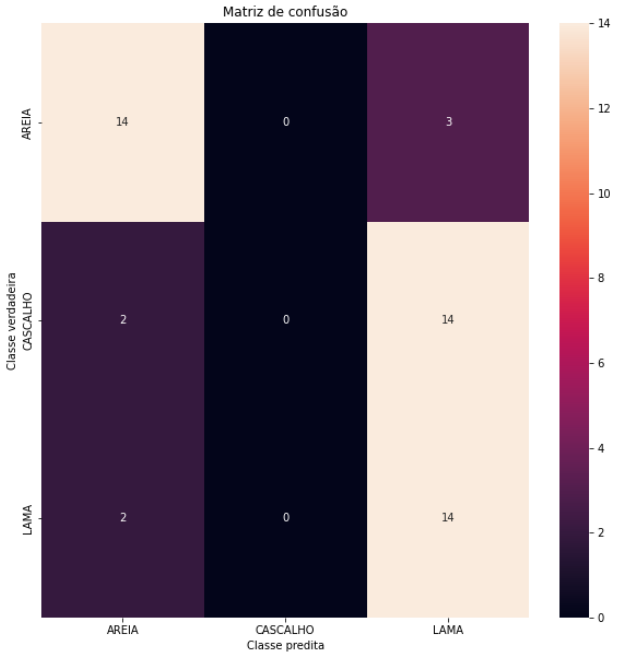
Para fins comparativos, nas figuras 32, 33, 34, 35 e 36 estão apresentadas as matrizes de confusão de cada K-fold atribuído para validação dos dados.

Figura 32 – Matriz de Confusão AdaBoost quando K-Fold = 1.



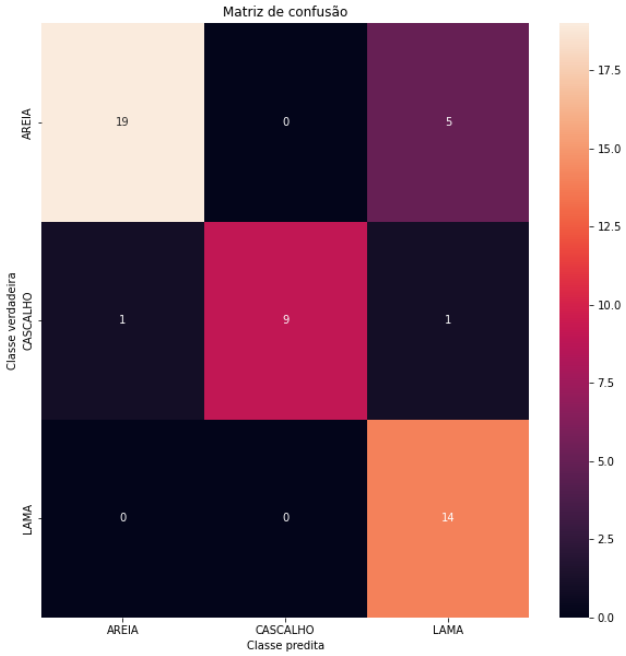
Fonte: Autor.

Figura 33 – Matriz de Confusão AdaBoost quando K-Fold = 2.



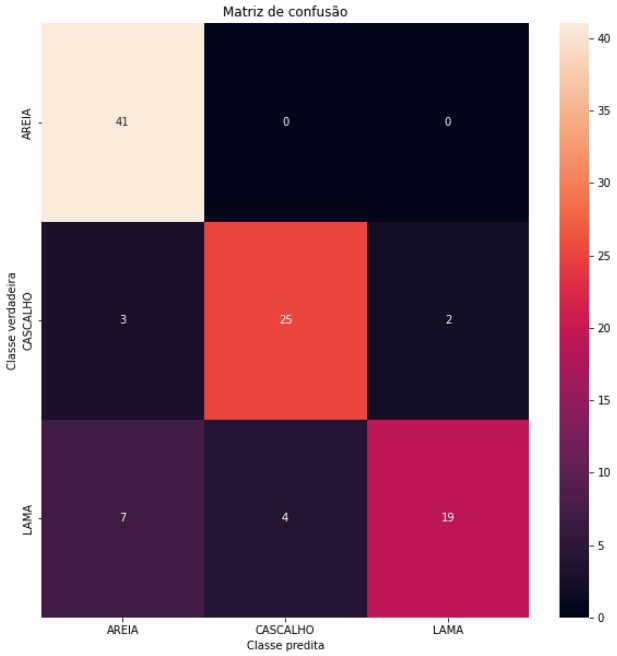
Fonte: Autor.

Figura 34 – Matriz de Confusão AdaBoost quando K-Fold = 3.



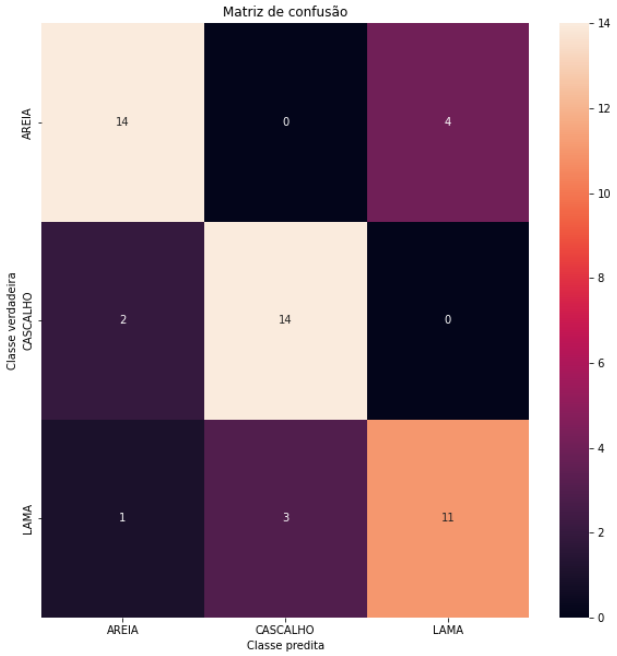
Fonte: Autor.

Figura 35 – Matriz de Confusão AdaBoost quando K-Fold = 4.



Fonte: Autor.

Figura 36 – Matriz de Confusão AdaBoost quando K-Fold = 5.



Fonte: Autor.

A Tabela 6 abaixo, apresenta os resultados do classificador SVM. O algoritmo foi implementado com a seguinte configuração: C = 1, kernel = rbf, degree = 3, gamma = scale. O resultado do SVM foi em média de 83,21% de acurácia, 84,11% de precisão, 83,21% para recall e 83,66% em F1-score. Os melhores valores de *Accuracy* encontrados no classificador SVM foram obtidos quando K = 3 com 83,22% de acurácia.

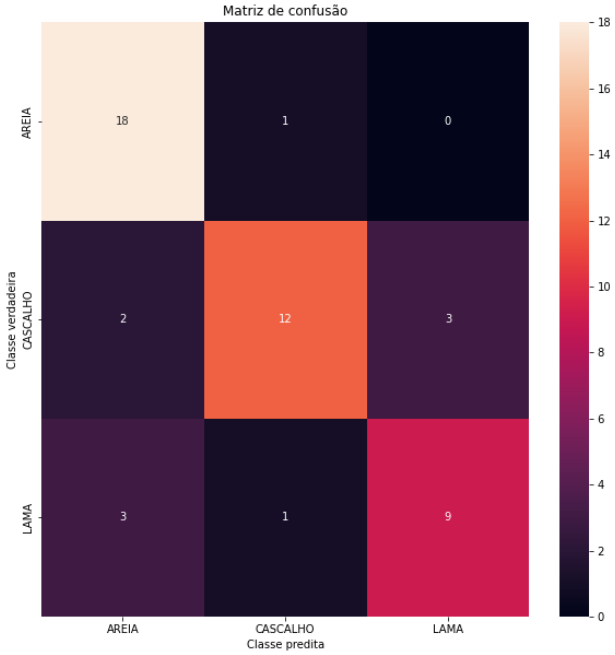
Tabela 6 – Resultados do Classificador SVM.

K-FOLD	**Accuracy**	**Precision**	**Recall**	**F1-Score**
K = 1	79,59%	79,98%	79,59%	79,79%
K = 2	85,71%	87,45%	85,71%	86,57%
K = 3	91,84%	93,00%	91,84%	92,42%
K = 4	75,25%	75,32%	75,25%	75,28%
K = 5	83,67%	84,80%	83,67%	84,23%
MÉDIA	**83,21%**	**84,11%**	**83,21%**	**83,66%**

Fonte: Autor.

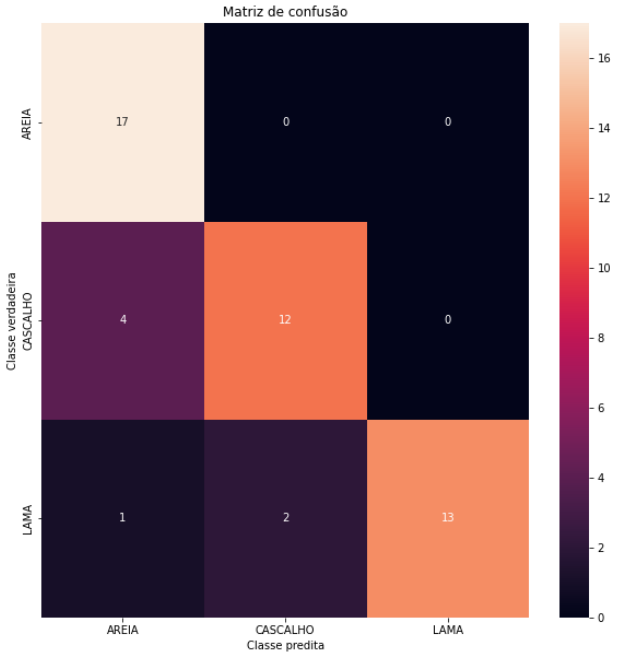
Para fins comparativos, nas figuras 37, 38, 39, 40 e 41 estão apresentadas as matrizes de confusão de cada K-fold atribuído para validação dos dados.

Figura 37 – Matriz de Confusão SVM quando K-Fold = 1.



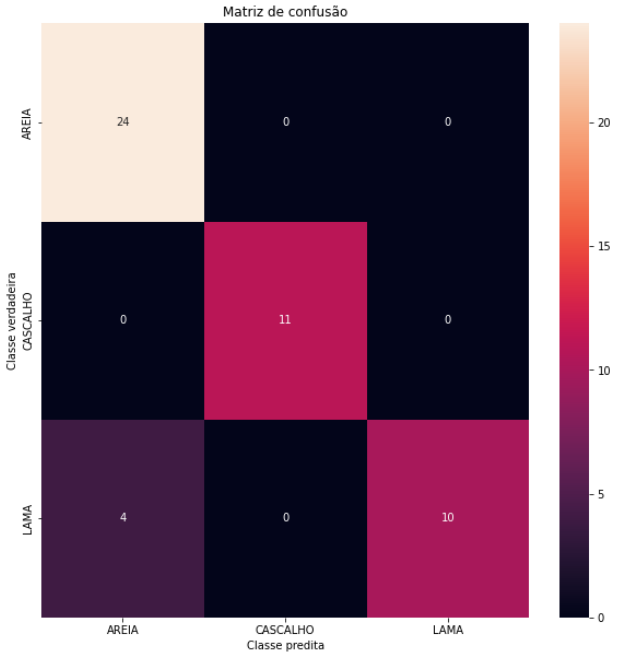
Fonte: Autor.

Figura 38 – Matriz de Confusão SVM quando K-Fold = 2.



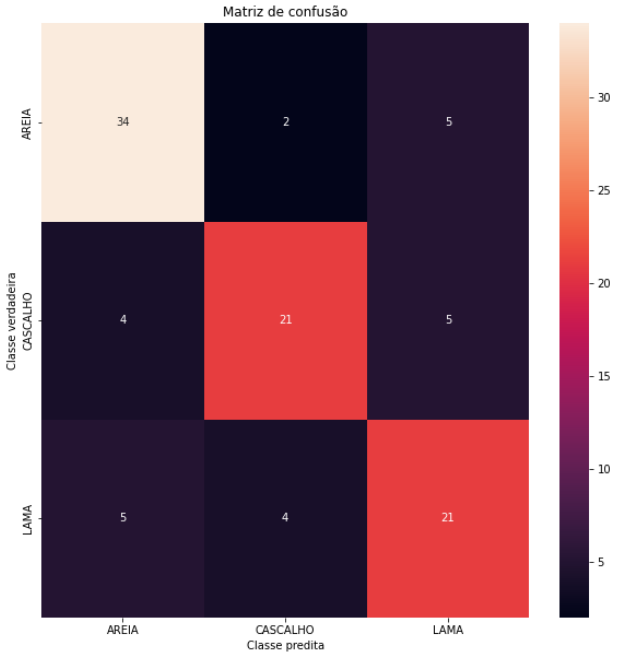
Fonte: Autor.

Figura 39 – Matriz de Confusão SVM quando K-Fold = 3.



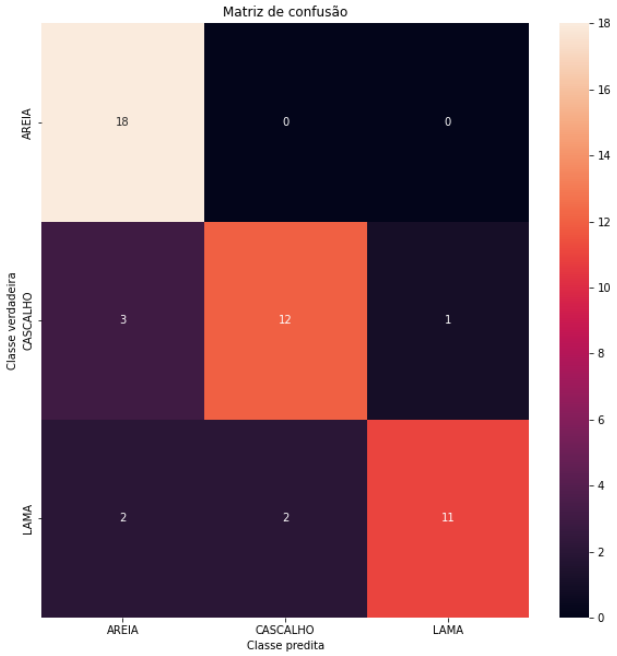
Fonte: Autor.

Figura 40 – Matriz de Confusão SVM quando K-Fold = 4.



Fonte: Autor.

Figura 41 – Matriz de Confusão SVM quando K-Fold = 5.



Fonte: Autor.

Os algoritmos de aprendizado de máquina têm uma importância significativa em diversas áreas do conhecimento, e neste estudo, esses algoritmos mostram que é possível encontrar padrões de sedimentos do fundo do mar. Apesar dos dados analisados possuírem aspectos peculiares, e pouca quantidade de dados, percebeu-se que esses resultados podem impulsionar ainda mais os avanços e inovações em tecnologia aplicada à geologia e a caracterização do fundo do mar, a fim de explorar o campo para uma possível perfuração de poços de petróleo na área da Margem Equatorial.

Os classificadores avaliados (*Multilayer Perceptron (MLP), Decision Trees, K-Nearest Neighbors (KNN), AdaBoost e Support Vector Machine (SVM)*) foram ajustados, configurados de forma específica para otimização do desempenho. Dessa forma, a comparação desses resultados destaca a eficácia da MLP em relação aos outros classificadores avaliados, apresentando a maior acurácia média.

Outra metodologia usou o classificador *Decision Trees* com dados de batimetria e retroespalhamento, para classificar quatro tipos de sedimentos na Baía de Santa Mônica. O resultado encontrado foi uma acurácia de 72% no mapeamento de fácies do fundo do mar (DARTNELL; GARDNER, 2004).

O trabalho de Cui et al. (2021), utilizou o método de aprendizagem profunda *Deep Belief Network* (DBN) para construir um modelo supervisionado de classificação de sedimentos no sul do Mar da Irlanda. Conseguiu 86,20% de acurácia na classificação de 10 tipos de sedimentos (areia levemente lamacenta, areia levemente cascalhada, lama de cascalho, areia lamacenta de cascalho, areia lamacenta, areia lamacenta, cascalho arenoso lamacento, areia, cascalho arenoso e lama arenosa).

Os algoritmos de aprendizado de máquina possuem diversos benefícios e aplicações, como, por exemplo, na eficiência na análise de grandes volumes de dados, sendo essencial ao lidar com dados coletados de áreas extensas do fundo do mar. Além disso, é interessante para a identificação de padrões complexos, com a capacidade de generalização de problema e na adaptação a ambientes variáveis, levando assim, a redução de erros humanos e o suporte à pesquisa científica e conservação ambiental. Dessa forma, uma classificação precisa dos sedimentos submarinos é essencial para a pesquisa científica, tendo a exploração de recursos naturais e a conservação ambiental para determinada finalidade. Na área da geologia, os algoritmos de aprendizado de máquina podem ajudar os cientistas a entender melhor os ecossistemas marinhos e a tomar decisões informadas sobre a gestão sustentável dos recursos marinhos.

Por fim, os algoritmos de aprendizado de máquina desempenham um papel fundamental na classificação de sedimentos do fundo do mar, proporcionando uma abordagem eficaz e precisa para analisar e compreender os ambientes submarinos complexos.

5 CONSIDERAÇÕES FINAIS

Classificar sedimentos marinhos através de algoritmos de aprendizado de máquina é uma boa alternativa por conta do custo envolvido e o tempo necessário para realizar a atividade. Aqui apresentou-se uma abordagem metodológica com cinco algoritmos distintos de aprendizado de máquina para avaliar a precisão de classificação dos sedimentos.

Foi observado que os algoritmos *Support Vector Machine e Multilayer Perceptron* apresentaram resultados maiores que 80% de acurácia, ou seja, eles conseguiram classificar corretamente mais de 80% dos dados separados para testes do modelo. Respectivamente a *Accuracy*, *Precision*, *Recall* e *F1_Score* encontrados no SVM foram: 83,21%, 84,11%, 83,21% e 83,66%. O algoritmo que teve o melhor desempenho foi o MLP com suas respectivas métricas: 83,99%, 84,56%, 83,99% e 84,27%.

Com esses resultados, observa-se que é possível criar um modelo de aprendizado de máquina para classificar automaticamente os sedimentos marinhos na área de estudo. Espera-se que esses algoritmos possam otimizar o mapeamento de fácies em toda a plataforma continental, subsidiando pesquisas futuras quanto a classificação das fácies sedimentares, auxiliando nas tomadas de decisão quanto ao ordenamento do uso espacial marítimo e na regulação das diversas atividades industriais na Margem Equatorial Brasileira.

Para trabalhos futuros, propõe-se utilizar outros dados além dos sedimentos superficiais do BNDO e das linhas sísmicas da Bacia da Margem Equatorial Brasileira, por exemplo imagens de satélites. Assim, seria possível melhorar o treinamento do classificador utilizando mais dados.

REFERÊNCIAS

ANP. *As áreas em oferta na Décima Primeira Rodada de Licitações*. 2022. Disponível em: <https://www.gov.br/anp/pt-br/rodadas-anp/rodadas-concluidas/concessao-de-blocos-exploratorios/11a-rodada-licitacoes-blocos/arquivos/seminarios/areas_em_oferta_r11.pdf>. Acesso em: 25 set. 2023. Citado na página 18.

AROSIO, R. et al. Fully convolutional neural networks applied to large-scale marine morphology mapping. *Frontiers in Marine Science*, Frontiers Media SA, v. 10, jul. 2023. ISSN 2296-7745. Disponível em: <http://dx.doi.org/10.3389/fmars.2023.1228867>. Citado na página 16.

AZEVEDO, L. P. de. Aplicação de redes neurais artificiais no processo de classificação de orquídeas do gênero cattleya. *Trabalho de Conclusão de Curso (Graduação em Sistemas de Informação) - Instituto Federal de Minas Gerais*, 2016. Citado 3 vezes nas páginas 21, 22 e 24.

BORCHARTT, T. B. Análise de imagens termográficas para classificação de alterações na mama. *Tese (Doutorado em Computação) - Universidade Federal Fluminense*, 2013. Citado na página 36.

BROWN, C. J. et al. Benthic habitat mapping: A review of progress towards improved understanding of the spatial ecology of the seafloor using acoustic techniques. *Estuarine, Coastal and Shelf Science*, Elsevier BV, v. 92, n. 3, p. 502–520, maio 2011. ISSN 0272-7714. Disponível em: <http://dx.doi.org/10.1016/j.ecss.2011.02.007>. Citado na página 16.

CHAVES, E. de L. Detecção de câncer de mama por meio de imagens infravermelhas utilizando redes neurais convolucionais. *Trabalho de Conclusão de Curso (Graduação em Ciência da Computação) - Universidade Federal de Uberlândia*, 2019. Citado na página 25.

CORTES, C.; VAPNIK, V. Support-vector networks. *Machine learning*, Springer, v. 20, n. 3, p. 273–297, 1995. Citado na página 28.

COSTA, A. H. M. da. Aplicação de redes neurais convolucionais na identificação de tatuadores. *Trabalho de Conclusão de Curso (Graduação em Engenharia Elétrica) - Universidade de Brasília*, 2018. Citado 2 vezes nas páginas 35 e 36.

CUI, X. et al. Deep learning model for seabed sediment classification based on fuzzy ranking feature optimization. *Marine Geology*, Elsevier BV, v. 432, p. 106390, fev. 2021. ISSN 0025-3227. Disponível em: <http://dx.doi.org/10.1016/j.margeo.2020.106390>. Citado 3 vezes nas páginas 15, 16 e 54.

DARTNELL, P.; GARDNER, J. V. Predicting seafloor facies from multibeam bathymetry and backscatter data. *Photogrammetric Engineering amp; Remote Sensing*, American Society for Photogrammetry and Remote Sensing, v. 70, n. 9, p. 1081–1091, set. 2004. ISSN 0099-1112. Disponível em: <http://dx.doi.org/10.14358/PERS.70.9.1081>. Citado 2 vezes nas páginas 15 e 54.

DIAS, M. F. R.; PASCUTTI, P. G.; SILVA, M. L. da. Aprendizado de máquina e suas aplicações em bioinformática. *Semioses*, v. 10, n. 1, p. 23–37, 2016. Citado na página 34.

DIESING, M. et al. Limitations of predicting substrate classes on a sedimentary complex but morphologically simple seabed. *Remote Sensing*, MDPI AG, v. 12, n. 20, p. 3398, out. 2020. ISSN 2072-4292. Disponível em: <http://dx.doi.org/10.3390/rs12203398>. Citado 2 vezes nas páginas 15 e 16.

FILHO, J. R. d. S. et al. Resizing the extension of the mesophotic "reefs" in the brazilian equatorial margin using bioclastic facies and seabed morphology. Research Square Platform LLC, ago. 2022. Disponível em: <http://dx.doi.org/10.21203/rs.3.rs-1927169/v1>. Citado 3 vezes nas páginas 18, 19 e 20.

FRANCKE, T.; LóPEZ-TARAZóN, J.; SCHRöDER, B. Estimation of suspended sediment concentration and yield using linear models, random forests and quantile regression forests. *Hydrological Processes*, Wiley, v. 22, n. 25, p. 4892–4904, ago. 2008. ISSN 1099-1085. Disponível em: <http://dx.doi.org/10.1002/hyp.7110>. Citado na página 15.

HASAN, R.; IERODIACONOU, D.; MONK, J. Evaluation of four supervised learning methods for benthic habitat mapping using backscatter from multi-beam sonar. *Remote Sensing*, MDPI AG, v. 4, n. 11, p. 3427–3443, nov. 2012. ISSN 2072-4292. Disponível em: <http://dx.doi.org/10.3390/rs4113427>. Citado 2 vezes nas páginas 15 e 16.

HAYKIN, S. *Redes Neurais - Princípios e Prática.* [S.l.]: Bookman, 2007. Citado 5 vezes nas páginas 21, 22, 23, 24 e 25.

HERBRICH, R. *Learning kernel classifiers: theory and algorithms.* [S.l.]: MIT press, 2001. Citado na página 30.

IERODIACONOU, D. et al. Comparison of automated classification techniques for predicting benthic biological communities using hydroacoustics and video observations. *Continental Shelf Research*, Elsevier BV, v. 31, n. 2, p. S28–S38, fev. 2011. ISSN 0278-4343. Disponível em: <http://dx.doi.org/10.1016/j.csr.2010.01.012>. Citado na página 15.

IZBICKI, R.; SANTOS, T. M. dos. *Aprendizado de máquina: uma abordagem estatística.* [S.l.]: Rafael Izbicki, 2020. Citado 5 vezes nas páginas 20, 21, 26, 27 e 29.

LARSONNEUR, C. *1977. La cartographie des dépôt meubles sur le plateau continental français: méthode mis.* 1977. Citado na página 31.

LE et al. Application of long short-term memory (lstm) neural network for flood forecasting. *Water*, MDPI AG, v. 11, n. 7, p. 1387, Jul 2019. ISSN 2073-4441. Disponível em: <http://dx.doi.org/10.3390/w11071387>. Citado na página 25.

LI, J.; TRAN, M.; SIWABESSY, J. Selecting optimal random forest predictive models: A case study on predicting the spatial distribution of seabed hardness. *PLOS ONE*, Public Library of Science (PLoS), v. 11, n. 2, p. e0149089, fev. 2016. ISSN 1932-6203. Disponível em: <http://dx.doi.org/10.1371/journal.pone.0149089>. Citado na página 15.

LORENA, A. C.; CARVALHO, A. C. de. Uma introdução às support vector machines. *Revista de Informática Teórica e Aplicada*, v. 14, n. 2, p. 43–67, 2007. Citado na página 30.

LUCIEER, V. et al. Do marine substrates 'look' and 'sound' the same? supervised classification of multibeam acoustic data using autonomous underwater vehicle images.

Estuarine, Coastal and Shelf Science, Elsevier BV, v. 117, p. 94–106, jan. 2013. ISSN 0272-7714. Disponível em: <http://dx.doi.org/10.1016/j.ecss.2012.11.001>. Citado na página 15.

MAYER, L. et al. The nippon foundation—gebco seabed 2030 project: The quest to see the world's oceans completely mapped by 2030. *Geosciences*, MDPI AG, v. 8, n. 2, p. 63, fev. 2018. ISSN 2076-3263. Disponível em: <http://dx.doi.org/10.3390/geosciences8020063>. Citado na página 15.

MILANI, E. J. et al. Petróleo na margem continental brasileira: geologia, exploração, resultados e perspectivas. *Revista Brasileira de Geofísica*, FapUNIFESP (SciELO), v. 18, n. 3, p. 352–396, 2000. ISSN 0102-261X. Disponível em: <http://dx.doi.org/10.1590/S0102-261X2000000300012>. Citado 2 vezes nas páginas 17 e 18.

MITCHELL, T. M. *Machine Learning*. New York: McGraw-Hill, 1997. ISBN 978-0-07-042807-2. Citado 2 vezes nas páginas 26 e 27.

MURTAGH, F. Multilayer perceptrons for classification and regression. *Neurocomputing*, v. 2, n. 5, p. 183–197, 1991. ISSN 0925-2312. Disponível em: <https://www.sciencedirect.com/science/article/pii/0925231291900235>. Citado na página 34.

NAQA, I. E.; MURPHY, M. J. What is machine learning? In: *machine learning in radiation oncology*. [S.l.]: Springer, 2015. p. 3–11. Citado na página 20.

PEDREGOSA, F. et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, v. 12, p. 2825–2830, 2011. Citado 2 vezes nas páginas 26 e 27.

QIN, X. et al. Optimizing the sediment classification of small side-scan sonar images based on deep learning. *IEEE Access*, Institute of Electrical and Electronics Engineers (IEEE), v. 9, p. 29416–29428, 2021. ISSN 2169-3536. Disponível em: <http://dx.doi.org/10.1109/access.2021.3052206>. Citado 2 vezes nas páginas 15 e 16.

SILVA, A. R. da; VELASQUEZ, L. H. *Uma visão geral sobre machine learning – Classificação*. 2020. Disponível em: <https://operdata.com.br/blog/uma-visao-geral-sobre-machine-learning/>. Acesso em: 15 dez. 2021. Citado 4 vezes nas páginas 21, 26, 27 e 28.

SILVA, I. N. da; SPATTI, D. H.; FLAUZINO, R. A. *Redes Neurais Artificiais Para Engenharia e Ciências Aplicadas*. [S.l.]: Artliber, 2016. Citado 4 vezes nas páginas 22, 23, 24 e 25.

APÊNDICE A – ARTIGO PDPETRO 2022

11º CONGRESSO BRASILEIRO DE PESQUISA E DESENVOLVIMENTO EM PETRÓLEO E GÁS

TÍTULO DO TRABALHO:

CLASSIFICAÇÃO DE SEDIMENTOS DA MARGEM EQUATORIAL BRASILEIRA UTILIZANDO ALGORITMOS DE APRENDIZADO DE MÁQUINA

AUTORES:

Luiz Eugênio Hoffmann Lopes[1], Allan Kardec Duailibe Barros Filho[1], Cleverson Guizan Silva[2], Diego de Oliveira Dantas[1], João Regis dos Santos Filho[2], Marta de Oliveira Barreiros[1].

INSTITUIÇÃO:

[1]Departamento de Engenharia Elétrica, Laboratório de Processamento de Informação Biológica (PIB), Universidade Federal do Maranhão (UFMA).
[2]Departamento de Geologia, LAGEMAR, Universidade Federal Fluminense (UFF).

CLASSIFICAÇÃO DE SEDIMENTOS DA MARGEM EQUATORIAL BRASILEIRA UTILIZANDO ALGORITMOS DE APRENDIZADO DE MÁQUINA

Resumo

Recentemente, importantes discussões a respeito da ocorrência de extensos recifes carbonáticos e do potencial de descobertas de hidrocarbonetos nas bacias do Pará-Maranhão e Foz do Amazonas, elevaram a importância do estudo da composição faciológica da plataforma continental. No entanto, amostrar sedimentos marinhos é uma operação de alto custo e muitas vezes demorada a depender da distância da costa e da profundidade da amostragem, sendo necessário o deslocamento da equipe e dos equipamentos através de embarcações de pesquisa especializadas. Tendo isto em vista, o objetivo principal deste trabalho foi integrando informações sedimentológicas de amostras de fundo e atributos sísmicos a partir de algoritmos de aprendizado de máquina para assim otimizar o mapeamento das fácies na plataforma continental da Margem Equatorial Brasileira. Os dados de sedimentos foram adquiridos do Banco Nacional de Dados Oceanográficos (BNDO), sendo classificados em Lama, Areia e Cascalho para o tamanho do grão e em Litoclástico (<30%), Litobioclástico (<50%), Biolitoclástico (<70%) e Bioclástico (>70%) para o teor de carbonato de cálcio. Os atributos sísmicos utilizados foram Coerência, Amplitude, Intensidade, Fase Instantânea e Frequência Instantânea. Foram desenvolvidos cinco algoritmos de aprendizado de máquina com a finalidade de relacionar os tipos de sedimento e teor de carbonatos aos atributos sísmicos. Os resultados de cada algoritmo são: *Support Vector Machine*: *Accuracy* 80,31%, *Precision* 81,21%, *Recall* 80,31% e *F1_Score* 80,01%. *K-Nearest Neighbors*: *Accuracy* 77,07%, *Precision* 80,55%, *Recall* 77,07% e *F1_Score* 75,94%. *Decision Trees*: *Accuracy* 85%, *Precision* 85,78%, *Recall* 85% e *F1_Score* 84,85%. *Multilayer Perceptron*: *Accuracy* 82,74%, *Precision* 82,74%, *Recall* 82,74% e *F1_Score* 82,49%. *Adaptive Boosting*: *Accuracy* 70,47%, *Precision* 72,25%, *Recall* 70,47% e *F1_Score* 69,75%. O melhor resultado obtido foi do algoritmo *Decision Trees* com 85% de acurácia. Com esses resultados preliminares, observa-se que é possível criar um modelo de aprendizado de máquina para classificar automaticamente os sedimentos marinhos, viabilizando e otimizando pesquisas futuras sobre as fácies carbonáticas e recifes mesofóticos na Margem Equatorial Brasileira.

Palavras-chave: Aprendizado de Máquina, Classificação, Margem Equatorial Brasileira, Sedimentos Marinhos.

Abstract

Recently, important discussions about the occurrence of extensive carbonate reefs and the potential for hydrocarbon discoveries in the Pará-Maranhão and Foz do Amazonas basins have raised the importance of studying the faciological composition of the continental shelf. However, sampling marine sediments is a high-cost and often time-consuming operation, depending on the distance from the coast and the depth of sampling, requiring the displacement of staff and equipment through specialized research vessels. With this in mind, the main objective of this work was to integrate sedimentological information from bottom samples and seismic attributes from machine learning algorithms to optimize the mapping of facies on the continental shelf of the Brazilian Equatorial Margin. Sediment data were acquired from the National Oceanographic Data Bank (BNDO), classified as Mud, Sand and Gravel for grain size and Lithoclastic (<30%), Lithobioclastic (<50%), Biolithoclastic (<70%) and Bioclastic (>70%) for calcium carbonate content. The seismic attributes used were Coherence, Amplitude, Intensity, Instantaneous Phase and Instantaneous Frequency. Five machine learning

algorithms were developed in order to relate sediment types and carbonate content to seismic attributes. The results of each algorithm are: Support Vector Machine: Accuracy 80.31%, Precision 81.21%, Recall 80.31% and F1_Score 80.01%. K-Nearest Neighbors: Accuracy 77.07%, Precision 80.55%, Recall 77.07% and F1_Score 75.94%. Decision Trees: Accuracy 85%, Precision 85.78%, Recall 85% and F1_Score 84.85%. Multilayer Perceptron: Accuracy 82.74%, Precision 82.74%, Recall 82.74% and F1_Score 82.49%. Adaptive Boosting: Accuracy 70.47%, Precision 72.25%, Recall 70.47% and F1_Score 69.75%. The best result was obtained from the Decision Trees algorithm with 85% accuracy. With these preliminary results, it is possible to create a machine learning model to automatically classify marine sediments, enabling and optimizing future research on carbonate facies and mesophotic reefs in the Brazilian Equatorial Margin.

Keywords: Classification, Brazilian Equatorial Margin, Machine Learning, Marine Sediments.

Introdução

A crescente demanda energética tem elevado a importância da Margem Equatorial Brasileira como uma nova fronteira para a extração de recursos não-renováveis como o petróleo e o gás natural, uma vez que a mesma tem semelhanças geológicas com a costa ocidental da África, que já é reconhecida mundialmente por suas reservas petrolíferas. Como as margens Brasileira e Africana são análogas, é possível ter as mesmas condições petrolíferas no Brasil (FILHO et al, 2020), conforme confirmadas pelas recentes descobertas de reservatórios nas bacias de Barreirinhas e Pará-Maranhão (PELLEGRINI e RIBEIRO, 2018). Ao todo, cerca de 18 campos produtores de petróleo já estão consolidados na Margem Equatorial, principalmente nas Bacias Potiguar e Ceará (ANP, 2020).

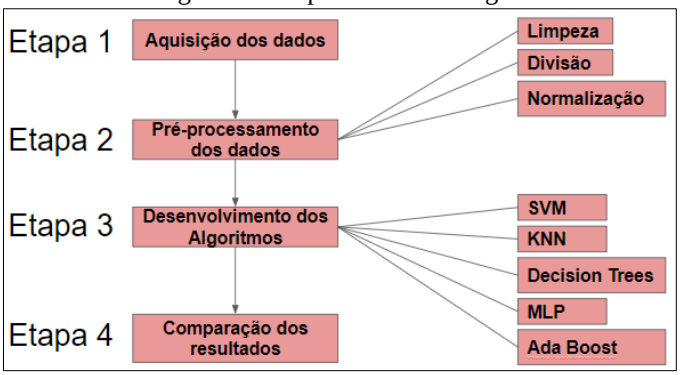
Além disso, os aproximadamente 9.650 km de litoral apresentam condições favoráveis para a implementação de fazendas eólicas offshore no Brasil, cujas políticas de regulação voltadas ao licenciamento, monitoramento e proteção ambiental ainda estão em discussão (GONZÁLEZ *et al.*, 2020). Nesse contexto, as medidas regulatórias, o planejamento estratégico e o uso do espaço marítimo passam pela necessidade de se refinar o reconhecimento do leito marinho onde serão instalados os aerogeradores. Apenas na Margem Equatorial estão em análise até o momento 31 áreas de Complexos Eólicos Offshore sobre a plataforma continental (IBAMA, 2022).

Como o Brasil tem uma extensa região costeira, e uma margem continental adjacente que por vezes ultrapassa as 200 milhas náuticas da sua Zona Econômica Exclusiva (ZEE), a realização da amostragem extensiva dos sedimentos marinhos não é eficiente ou operacionalmente viável para o ideal reconhecimento das fácies sedimentares da plataforma continental. Este trabalho utilizou dados públicos de sedimento e sísmica multicanal 2D disponíveis na região da Margem Equatorial Brasileira para prever qual é a sua classe de sedimento com o auxílio de aprendizado de máquina.

Metodologia

Este trabalho foi dividido em 4 etapas: Aquisição dos dados, Pré-processamento dos dados, Desenvolvimento dos algoritmos e Comparação dos resultados. A figura 1 abaixo, mostra a sequência de etapas realizadas.

Figura 1 – Etapas da Metodologia.



Fonte: Autor.

Aquisição dos dados

Ao todo, foram reunidos 5.515 dados de sedimentos superficiais provenientes do Banco Nacional de Dados Oceanográficos (BNDO) e do Banco de Dados da Indústria do Petróleo (BAMPETRO), apresentando duas informações básicas: natureza do fundo oceânico (feita por descrição visual) ou através de dados sedimentológicos processados em laboratório. Os teores de tamanho de grão foram classificados segundo a classificação ternária de SHEPARD (1954) em três classes principais, sendo: Lama, Areia e Cascalho, e o teor de carbonato de cálcio em quatro classes, sendo: Litoclástico ($CaCO_3$<30%), Litobioclástico (30%<$CaCO_3$<50%), Litobioclástico (50%<$CaCO_3$<70%) e Litoclástico ($CaCO_3$>70%) (LARSONNEUR, 1977 adaptado por DIAS, 1996).

Os dados de atributos sísmicos consistem em resultados da interpretação e processamento de linhas públicas de sísmica 2D multicanal extraídos no software Paleoscan™. Ao todo foram utilizadas 384 linhas sísmicas. Os primeiros 500 milissegundos de cada linha sísmica tiveram seu fundo do mar interpretado manualmente para a extração de cinco atributos sísmicos: amplitude, intensidade de reflexão, fase, frequência e coerência.

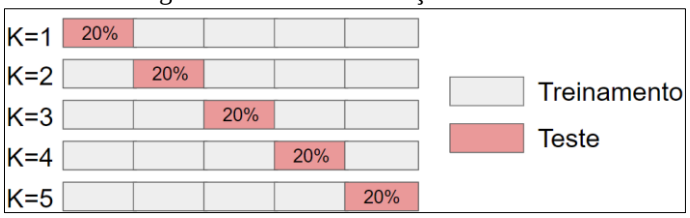
O resultado da etapa 1 foi um arquivo ".csv" com as 5.515 linhas referentes aos pontos de sedimento e 106 colunas dos mais diversos tipos de parâmetros sedimentológicos. Na etapa seguinte, foi realizado o pré-processamento desses dados para serem utilizados na entrada dos algoritmos.

Pré-processamento dos dados

No pré-processamento dos dados foi feita a limpeza, a divisão e a normalização. Inicialmente são excluídos os dados inconsistentes ou nulos. Depois são separadas apenas as colunas de interesse: Classe_Shepard (Areia, Cascalho, Lama), a latitude e a longitude do sedimento, e os cinco atributos sísmicos associados. O resultado foi um arquivo “.csv” com 8 colunas, 101 linhas de areia, 74 linhas de cascalho e 73 linhas de lama.

A Figura 2 abaixo, mostra como foi realizada a Validação Cruzada dos dados. Se K for igual a 5, significa que os dados são divididos em 5 partes (20% dos dados), ou seja, o algoritmo executa 5 vezes, onde na primeira execução a parte 1 será destinada para teste e o restante para treinamento. Já na segunda execução, a parte 2 dos dados será destinada para teste e o restante para treinamento, assim sucessivamente até K = 5.

Figura 2 – Divisão Validação Cruzada.



Fonte: Autor.

A técnica de Validação Cruzada é utilizada para avaliar a capacidade de generalização de um modelo, com ela é possível realizar o teste do algoritmo em todos os dados.

Desenvolvimento dos algoritmos

Neste trabalho foram desenvolvidos cinco algoritmos de aprendizado de máquina, descritos a seguir:

i- *Support Vector Machine* (SVM), ou Máquinas de Vetores de Suporte, constituem uma técnica de Aprendizado de Máquina embasada na teoria de aprendizado estatístico, desenvolvida por VAPNIK (1995). Tem uma série de princípios para obter classificadores com boa generalização, ou seja, boa capacidade de prever corretamente a classe de novos dados do mesmo domínio em que o aprendizado ocorreu (VALDECY, 2013).

O fundamento básico do funcionamento de uma SVM é a construção de um hiperplano (objeto matemático), o qual pode possuir uma dimensão R^n, possibilitando a separação dos dados em classes distintas. Essa separação tem abordagem baseada na Teoria de Aprendizagem Estatística, utilizando a otimização matemática, para promover a busca de um hiperplano ótimo, adotando os mínimos quadrados por meio da minimização estrutural de risco (BARRETO et al, 2019).

ii- *K-Nearest Neighbors* (KNN) é um método de classificação muito popular em mineração de dados, regressão e imputação de valor ausente, por causa de sua implementação simples e desempenho de classificação significativo. Realiza tarefas de classificação calculando primeiro a distância entre a amostra de teste e todas as outras amostras de treinamento para obter seus vizinhos mais próximos e, em seguida, classifica as amostras de teste com rótulos pela regra da maioria nos rótulos de vizinhos mais próximos selecionados (ZHANG et al, 2018).

iii- *Decision Trees*, ou Árvores de decisão, são modelos de processos biológicos e cognitivos, são formas simples e eficazes de representar o conhecimento. Para construção de uma *Decision Trees*, seleciona-se inicialmente um atributo de particionamento e divide o conjunto de exemplos, criando um ramo para cada valor deste atributo. A cada ramo é criado um nodo, e novamente selecionado um novo atributo de particionamento para o subconjunto de exemplos atualizando o nodo e, assim, sucessivamente. Tem como objetivo separar as classes em partições diferentes (GARCIA, 2003).

iv- *Multilayer Perceptron* (MLP) é uma rede neural artificial com uma ou mais camadas ocultas e um número indeterminado de neurônios, é uma derivação do *Perceptron* simples. O número de camadas ocultas em um *Perceptron* multicamadas e o número de neurônios em cada camada pode variar dependendo do problema. Em geral, mais neurônios oferecem maior sensibilidade ao problema a ser resolvido, mas também maior risco de *overfit* (MURTAGH, 1991).

v- *Adaptive Boosting*, também conhecido como *AdaBoost*, pode ser utilizado para aumentar a performance de outros algoritmos de aprendizagem. É adaptável no sentido de que as classificações subsequentes feitas são ajustadas a favor das instâncias classificadas negativamente por classificações anteriores. É sensível ao ruído nos dados e casos isolados. Os principais diferenciais desse algoritmo são: (i) requer baixo valor computacional, e (ii) a margem das fronteiras de decisão entre as classes é mais precisa que outros algoritmos predecessores (CHAVES, 2011).

Comparação dos resultados

As métricas mais utilizadas em aprendizado de máquina para avaliação da qualidade dos classificadores são: *Accuracy*, *Precision*, *Recall* e *F1-Score*. Onde: *Accuracy* – indica uma performance geral do modelo, ou seja, dentre todas as classificações, quantas o modelo classificou corretamente; *Precision* – encontra os Falsos Negativos; *Recall* – encontra os Falsos Positivos; *F1-Score* – é a média harmônica entre *Precision* e *Recall*.

Para realizar o trabalho, foi utilizada a linguagem de programação *python* e a biblioteca *sklearn* no ambiente *Google Colab Pro*, que fornece o seguinte *hardware*: GPU Tesla P100-PCIE-16GB, Intel Xeon @ 2x 2 GHz CPU, 13.021 MiB RAM, 70 GB HD e *Ubuntu* 18.04 *Linux Kernel* 4.19.104.

Resultados e Discussão

Foram realizados testes com os dados de latitude, longitude do sedimento e os cinco atributos sísmicos: amplitude, intensidade de reflexão, fase, frequência e coerência. Mas os melhores resultados foram encontrados utilizando latitude, longitude e apenas três dos atributos sísmicos: amplitude, fase e frequência. Os melhores resultados obtidos são apresentados nas tabelas a seguir.

Na Tabela 1 abaixo, são apresentados os resultados dos classificadores MLP, KNN e SVM. O MLP tem a seguinte configuração: 50 camadas ocultas, 100 neurônios por camadas, taxa de aprendizado 0,000001, função de ativação RELU e otimização ADAM. O algoritmo KNN tem como configuração o valor de K = 3, ou seja, os três vizinhos mais próximos são utilizados para classificação. O SVM tem a seguinte configuração: C = 1, kernel = rbf, degree = 3, gamma = scale.

Tabela 1 – Resultados dos Classificadores MLP, KNN e SVM.

K-FOLD	**MLP**				**KNN**				**SVM**			
	Accuracy	***Precision***	***Recall***	***F1-Score***	***Accuracy***	***Precision***	***Recall***	***F1-Score***	***Accuracy***	***Precision***	***Recall***	***F1-Score***
K = 1	75,51%	75,25%	75,51%	74,74%	79,59%	83,67%	79,59%	78,72%	77,55%	80,88%	77,55%	76,80%
K = 2	87,75%	87,52%	87,75%	87,58%	81,63%	86,87%	81,63%	80,77%	87,75%	87,74%	87,75%	87,21%
K = 3	77,55%	78,08%	77,55%	77,50%	71,42%	72,99%	71,42%	68,67%	73,46%	74,39%	73,46%	73,34%
K = 4	85,14%	85,10%	85,14%	85,03%	77,22%	80,56%	77,22%	76,67%	81,18%	81,31%	81,18%	81,09%
K = 5	87,75%	87,77%	87,75%	87,59%	75,51%	78,65%	75,51%	74,84%	81,63%	81,74%	81,63%	81,63%
MÉDIA	82,74%	82,74%	82,74%	82,49%	77,07%	80,55%	77,07%	75,94%	80,31%	81,21%	80,31%	80,01%

Fonte: Autor.

Na Tabela 2 abaixo, são apresentados os resultados dos classificadores *Decision Trees* e *AdaBoost*. Suas implementações foram feitas com as configurações padrões, conforme encontrada na biblioteca *sklearn*.

Tabela 2 – Resultados dos Classificadores *Decision Trees* e *AdaBoost*.

K-FOLD	Decision Trees				AdaBoost			
	Accuracy	**Precision**	**Recall**	**F1-Score**	**Accuracy**	**Precision**	**Recall**	**F1-Score**
K = 1	91,83%	92,34%	91,83%	91,88%	89,79%	91,29%	89,79%	89,90%
K = 2	85,71%	86,88%	85,71%	85,94%	69,38%	69,22%	69,38%	67,64%
K = 3	75,51%	76,84%	75,51%	74,84%	73,46%	75,32%	73,46%	74,06%
K = 4	82,17%	82,85%	82,17%	81,86%	56,43%	60,75%	56,43%	55,82%
K = 5	89,79%	89,98%	89,79%	89,71%	63,26%	64,65%	63,26%	61,34%
MÉDIA	85,00%	85,78%	85,00%	84,85%	70,47%	72,25%	70,47%	69,75%

Fonte: Autor.

Conclusões

Neste trabalho foram desenvolvidos cinco algoritmos para classificação dos sedimentos marinhos da Margem Equatorial Brasileira: Multilayer Perceptron, K-Nearest Neighbors, Decision Trees, Adaptive Boosting e Support Vector Machine. Os dois melhores resultados encontrados foram no algoritmo MLP com *Accuracy* 82,74%, *Precision* 82,74%, *Recall* 82,74% e *F1_Score* 82,49% e no algoritmo *Decision Trees* com *Accuracy* 85%, *Precision* 85,78%, *Recall* 85% e *F1_Score* 84,85%.

Com esses resultados, observa-se que é possível criar um modelo de aprendizado de máquina para classificar automaticamente os sedimentos marinhos na área de estudo. Espera-se que esses algoritmos possam otimizar o mapeamento de fácies em toda a plataforma continental, subsidiando pesquisas futuras quanto a classificação das fácies sedimentares, auxiliando nas tomadas de decisão quanto ao ordenamento do uso espacial marítimo e na regulação das diversas atividades industriais na Margem Equatorial Brasileira.

Agradecimentos

Agradecemos ao apoio financeiro da Agência Nacional do Petróleo, Gás Natural e Biocombustíveis (ANP) e da Financiadora de Estudos e Projetos (FINEP), por meio do Programa de Recursos Humanos da ANP para o Setor Petróleo e Gás - PRH/ANP (PRH 54.1 - UFMA). Também agradecemos a Agência Nacional do Petróleo, Gás Natural e Biocombustíveis (ANP) e a TGS pelo fornecimento dos dados de sísmica multicanal.

Referências Bibliográficas

ANP. **BDEP: Banco de Dados de Exploração e Produção**. Disponível em: <http://www.anp.gov.br/publicacoes/folderes/2416-bdep-banco-de-dados-de-exploracao-e-producao>.

Acesso em: 7 maio. 2022.

BARRETO, L. C.; CARVALHO, F. O.; SANTOS, B. C. V.; SOARES, A. P. M. R.; SILVA, G. D.. Aplicação de Support Vector Machine na Detecção de Falhas em Sensores de uma Coluna Debutanizadora. XIII Congresso Brasileiro de Engenharia Química em Iniciação Científica, **Blucher Chemical Engineering Proceedings**, Volume 1, 2019, Pages 3003-3009, ISSN 2359-1757. Disponível em: http://dx.doi.org/10.1016/cobecic2019-SOCP31.

CHAVES, B. B.. **Estudo do algoritmo AdaBoost de aprendizagem de máquina aplicado a sensores e sistemas embarcados**. 2011. Dissertação (Mestrado em Engenharia de Controle e Automação Mecânica) - Escola Politécnica, Universidade de São Paulo, São Paulo, 2011.

FILHO, A. K. D. B.; CARMONA, R. G.; ZALÁN, P. V.. Nota técnica sobre a margem equatorial brasileira. **Um novo "Pré-sal" no arco norte do território brasileiro?.** 2020.

GARCIA, S. C.. **O uso de árvores de decisão na descoberta de conhecimento na área da saúde**. 2003. Dissertação (Mestrado em Ciência da computação) - Universidade Federal do Rio Grande do Sul.

GONZÁLEZ, M. O. A.; SANTISO, A. M.; MELO, D. C. DE; VASCONCELOS, R. M. DE. Regulation for offshore wind power development in Brazil. **Energy Policy**, v. 145, n. June 2019, 2020.

IBAMA. **Mapas de projetos em licenciamento - Complexos Eólicos Offshore**. Disponível em: <https://www.ibama.gov.br/laf/consultas/mapas-de-projetos-em-licenciamento-complexos-eolicos-offshore>. Acesso em: 10 mar. 2022.

LARSONNEUR, C.. 1977. La cartographie des dépôt meubles sur le plateau continental français: méthode mise au point et utilisée en Manche. **J. Rech. Ocean**. 2, 34–39.

MURTAGH, F.. Multilayer perceptrons for classification and regression. **Neurocomputing**. Volume 2. Issues 5–6. 1991. Pages 183-197. ISSN 0925-2312. Disponível em: https://doi.org/10.1016/0925-2312(91)90023-5.

PELLEGRINI, B. S.; RIBEIRO, H. J. P. S.. Exploratory plays of Pará-Maranhão and Barreirinhas basins in deep and ultra-deep waters, Brazilian Equatorial Margin. **Brazilian Journal of Geology**, v. 48, n. 3, p. 485–502, 2018.

SHEPARD, F. P.. 1954. Nomenclature Based on Sand-silt-clay Ratios. ***SEPM J. Sediment****. Res.* doi:10.1306/d4269774-2b26-11d7-8648000102c1865d.

VALDECY, V. J.. **Classificação de dados usando técnicas de datamining e aprendizado de máquina**. 2013. Trabalho de Conclusão de Curso (Bacharelado em Ciência da Computação) - Universidade Federal do Maranhão.

VAPNIK, V. N., 1995. The nature of statistical learning theory. **Springer-Verlag New York**, Inc., New York, NY, USA. ISBN 0-387-94559-8.

ZHANG, S.; LI, X.; ZONG, M.; ZHU, X.; WANG, R.. Efficient kNN Classification With Different Numbers of Nearest Neighbors. 2018. **IEEE Transactions on Neural Networks and Learning Systems**, vol. 29, no. 5, pp. 1774-1785. Disponível em: https://doi.org/10.1109/TNNLS.2017.2673241.