CEFET/RJ
DIRETORIA DE PESQUISA E PÓS-GRADUAÇÃO

APLICAÇÃO DE TÉCNICAS DE APRENDIZADO DE MÁQUINA NA DETERMINAÇÃO DE ESTOQUE DE CARBONO NO SOLO.

Alexandre Emilio M. Pardelinha

Dissertação de Mestrado apresentada ao Programa de Pós-graduação em Ciência da Computação, Centro Federal de Educação Tecnológica Celso Suckow da Fonseca CEFET/RJ, como parte dos requisitos necessários à obtenção do grau de mestre.

Orientadores:
Diego Nunes Brandão
Laura Silva de Assis

Rio de Janeiro,

Fevereiro de 2025

Aplicação de Técnicas de Aprendizado de Máquina na Determinação de Estoque de Carbono no Solo.

Dissertação de Mestrado em Ciência da Computação, Centro Federal de Educação Tecnológica Celso Suckow da Fonseca, CEFET/RJ.

Alexandre Emilio M. Pardelinha

Aprovada por:

gov.br Documento assinado digitalmente
DIEGO NUNES BRANDAO
Data: 11/02/2025 15:22:36-0300
Verifique em https://validar.iti.gov.br

Presidente, Prof. Diego N. Brandão, D.Sc. (orientador)

LAURA SILVA DE ASSIS:06571252619
Assinado de forma digital por LAURA SILVA DE ASSIS:06571252619
Dados: 2025.02.11 07:01:45 -03'00'

Laura Silva de Assis, D.Sc. (coorientadora)

gov.br Documento assinado digitalmente
JORGE DE ABREU SOARES
Data: 10/02/2025 22:16:54-0300
Verifique em https://validar.iti.gov.br

Jorge de Abreu Soares, D.Sc. - CEFET/RJ

gov.br Documento assinado digitalmente
MARCOS BACIS CEDDIA
Data: 11/02/2025 09:34:22-0300
Verifique em https://validar.iti.gov.br

Marcos Bacis Ceddia, D.Sc. - UFRRJ

Rio de Janeiro,

Fevereiro de 2025

Ficha catalográfica elaborada pela Biblioteca Central do CEFET/RJ

P226 Pardelinha, Alexandre Emilio M.
Aplicação de técnicas de aprendizado de máquina na determinação de estoques de carbono no solo / Alexandre Emilio M. Pardelinha – 2025.
72f : il. (algumas color.) , enc.

Dissertação (Mestrado). Centro Federal de Educação Tecnológica Celso Suckow da Fonseca, 2025.
Bibliografia : f. 56-72.
Orientador: Diego Nunes Brandão.
Orientador: Laura Silva de Assis.

1. Aprendizado de máquina. 2. Carbono. 3. Solos. 4. Sequestro de carbono. I. Brandão, Diego Nunes (Orient.). II. Assis, Laura Silva de (Orient.). III. Título.

CDD 006.31

Elaborada pelo bibliotecário Leandro Mota de Menezes – CRB/7 nº 5281

DEDICATÓRIA

Dedico este trabalho ao meu filho Arthur da Silva Pardelinha, o meu alicerce para superar os desafios.

AGRADECIMENTOS

Primeiramente, agradeço à minha esposa Anne Pardelinha e ao meu filho Arthur Pardelinha, que foram essenciais para que eu pudesse superar os desafios e chegar até aqui.

Agradeço aos meus pais, Carlos Alberto Pardelinha e Sílvia C. M. Pardelinha, pelo incentivo.

Obrigado, também, aos meus orientadores Prof. Diego Nunes Brandão e Prof.ª Laura Assis, pelas idéias, orientações e paciência.

Aos meus amigos de trabalho no IBGE, Andrea Borges e Raquel Rose pelo incentivo da realização deste Mestrado.

Aos amigos que conquistei durante esta jornada no Cefet, que me ajudaram e incentivaram, em especial, Roberto Gervasio, Lucas Pinheiro e Augusto José da Fonseca.

RESUMO

Aplicação de Técnicas de Aprendizado de Máquina na Determinação de Estoque de Carbono no Solo.

Alexandre Emilio M. Pardelinha

Orientadores:
Diego Nunes Brandão
Laura Silva de Assis

Resumo da Dissertação submetida ao Programa de Pós-graduação em Ciência da Computação do Centro Federal de Educação Tecnológica Celso Suckow da Fonseca CEFET/RJ como parte dos requisitos necessários à obtenção do grau de mestre.

Os solos representam o mais significativo estoque de carbono orgânico (SOCS) nos ecossistemas terrestres, sublinhando a importância crítica de estimar com acurácia o carbono orgânico do solo para garantir a preservação das funções do solo e a mitigação das alterações climáticas globais. Este estudo emprega uma metodologia baseada em dados para estimar os estoques de carbono em solos brasileiros, comparando técnicas de aprendizado de máquina com diversas estratégias de otimização de hiperparâmetros. Após a seleção de atributos pelos métodos de Correlação de Pearson e *Wrapper*, o conjunto de dados referente à Amazônia consistiu em 479 observações de SOCS e 5 covariáveis de solo. Os desempenhos do modelo foram avaliados usando um procedimento de validação cruzada de 5 vezes. Os melhores desempenhos foram obtidos usando o algoritmo *Random Forest* (RMSE=0,938 com a técnica de otimização *Random Search* e MAE=0,706 com a técnica de otimização *Successive Halving*). Quase todos os resultados dos desempenhos nas três técnicas de otimização de hiperparâmetros apresentaram maior acurácia do que os desempenhos sem utilizar a estratégia de otimização. Os resultados demonstram o papel fundamental da seleção e processamento de dados, juntamente com a otimização de hiperparâmetros, na resolução deste problema, resultando em melhorias notáveis nas análises do erro médio absoluto (MAE) e da raiz do erro quadrático médio (RMSE).

Palavras-chave: Estoque de Carbono, Aprendizado de Máquina, Otimização de Hiperparâmetros

Rio de Janeiro,

Fevereiro de 2025

ABSTRACT

Aplicação de Técnicas de Aprendizado de Máquina na Determinação de Estoque de Carbono no Solo.

Alexandre Emilio M. Pardelinha

Advisors:
Diego Nunes Brandão
Laura Silva de Assis

Abstract of dissertation submitted to Programa de Pós-graduação em Ciência da Computação - Centro Federal de Educação Tecnológica Celso Suckow da Fonseca CEFET/RJ as partial fulfillment of the requirements for the degree of master.

Soils represent the most significant stock of organic carbon (SOCS) in terrestrial ecosystems, underscoring the critical importance of accurately estimating soil organic carbon to ensure the preservation of soil functions and the mitigation of global climate change. This study employs a data-driven methodology to estimate carbon stocks in Brazilian soils, comparing machine learning techniques with various hyperparameter optimization strategies. After attribute selection using the Pearson Correlation and *Wrapper* methods, the dataset for the Amazon consisted of 479 SOCS observations and 5 soil covariates. Model performances were evaluated using a 5-fold cross-validation procedure. The best performances were obtained using the *Random Forest* algorithm (RMSE=0.938 with the *Random Search* optimization technique and MAE=0.706 with the *Successive Halving* optimization technique). Almost all performance results in the three hyperparameter optimization techniques showed greater accuracy than the performances without using the optimization strategy. The results demonstrate the fundamental role of data selection and processing and hyperparameter optimization in solving this problem, resulting in notable improvements in mean absolute error (MAE) and root mean square error (RMSE) analyses.

Key-words: Carbon Stock, Machine Learning, Hyperparameter Optimization

Rio de Janeiro,

Fevereiro de 2025

Sumário

I Introdução **9**

II Referencial Teórico **11**

II.1 Ciclo global do carbono 11

II.2 Mapeamento de solos: tradicional e digital 12

II.3 Aprendizado de Máquina 14

II.3.1 Árvores de decisão 16

II.3.2 Floresta Aleatória 17

II.3.3 Gradient Boosting 17

II.4 Validação Cruzada *K-fold* 19

II.5 Seleção de Atributos 19

II.5.1 Correlação de Pearson 20

II.5.2 Métodos *Wrapper* 21

II.6 Otimização de Hiperparâmetros 21

II.7 Avaliação da Acurácia dos Modelos 23

III Trabalhos Relacionados **24**

III.1 Base de dados e String de Busca 24

III.2 Origem dos dados dos solos 25

III.3 Métodos de AM mais comumente usados nas ciências do solo 25

III.4 Métricas de desempenho mais comumente usadas nas ciências do solo 26

III.5 Frequência dos artigos referente aos anos encontrados no *Scopus* 27

III.6 Regressão ou classificação: o mais comumente usado nas ciências do solo 27

III.7 Modelo de AM que obteve melhor resultado 28

III.8 Número de estimadores usados no modelo *Random Forest* 28

III.9 Método de validação cruzada (*K-fold*) 29

III.10 Levantamentos dos dados de solos utilizados 29

III.11 Interpretabilidade dos modelos de AM 31

III.12 Os artigos avaliados 33

III.12.1Artigos que usaram RF 33
III.12.2Artigos que não usaram RF 36
III.13Artigo referência para comparação dos resultados 38

IV Metodologia 39
IV.1 Fluxo de Trabalho 39
IV.2 Os hiperparâmetros definidos 40
IV.2.1 Métodos *Wrapper* 40
IV.2.2 Otimização por hiperparâmetros 41
IV.3 Determinação do Estoque de Carbono Orgânico do solo 42
IV.4 Base de Dados 42
IV.5 Análise descritiva dos dados 44

V Resultados 47
V.1 Resultados da seleção de atributos 47
V.1.1 Correlação de Pearson 47
V.1.2 Métodos *Wrapper* 49
V.1.3 Observações referentes aos resultados da seleção de atributos 50
V.2 Determinação do número de iterações ideal 50
V.3 Comparação com os resultados do artigo de Ceddia et al. [2016] 51
V.4 Resultados da estimação de SOCS com Otimização de Hiperparâmetros 52

VI Considerações Finais 54

Referências Bibliográficas 56

Capítulo I Introdução

O solo desempenha um papel crucial na sustentabilidade da sociedade e dos ecossistemas, oferecendo uma gama de funções vitais. Ele também oferece uma ampla gama de serviços essenciais para a vida na Terra. Além de ser a base da produção de alimentos, o solo regula o clima, armazenando carbono e mitigando os efeitos das mudanças climáticas. Ele também contribui para a regulação do ciclo hidrológico, armazenando água e liberando-a gradualmente, o que ajuda a prevenir secas e inundações. Além disso, o solo abriga uma grande diversidade de organismos, contribuindo para a conservação da biodiversidade. [Ferreira et al., 2023].

Apesar de sua importância vital, os solos em todo o mundo enfrentam desafios crescentes, incluindo degradação devido à erosão, contaminação, urbanização desordenada e práticas agrícolas insustentáveis. Assim, compreender e preservar a saúde do solo é essencial para garantir a segurança alimentar, a sustentabilidade ambiental e a estabilidade climática em nosso planeta [Haddad et al., 2018]. Em relação a estabilidade climática, o solo tem um papel importante, pois atua como um vasto reservatório de carbono, desempenhando um papel crítico na regulação das concentrações de dióxido de carbono (CO_2) na atmosfera e, consequentemente, na mitigação das mudanças climáticas.

A compreensão da quantidade de Estoque de Carbono Orgânico no Solo (SOCS), fundamental para a regulação das mudanças climáticas, tem impulsionado a busca por técnicas avançadas capazes de determinar essa característica [Ye et al., 2021; Szatmari et al., 2023]. Dentre essas técnicas, destacam-se os algoritmos de Aprendizado de Máquina (AM), que vêm ganhando destaque nos últimos anos. Esse avanço está alinhado ao crescimento expressivo das atividades de mapeamento de solos, motivado pela crescente demanda por informações quantitativas e espaciais sobre o solo [Haddad et al., 2018; Wadoux et al., 2020a].

Os algoritmos de Aprendizado de Máquina (AM) destacam-se por sua capacidade de operar sem as limitações de suposições estatísticas rígidas, apresentando maior acurácia em comparação aos modelos estatísticos convencionais, especialmente ao lidar com grandes volumes de dados multicolineares. Contudo, muitos estudos têm aplicado AM como ‘caixas-pretas‘ [Wadoux et al., 2020a], frequentemente ignorando a importância da qualidade e do tratamento dos dados utilizados no treinamento desses modelos [Kumar et al., 2023a].

Em uma metodologia para determinação do estoque de carbono orgânico no solo (SOCS), a

otimização de hiperparâmetros é uma etapa fundamental para melhorar a acurácia e a confiabilidade dos modelos, características essenciais para o monitoramento ambiental e a mitigação das mudanças climáticas [Japa et al., 2023]. Este trabalho busca responder qual o impacto da otimização de hiperparâmetros na melhoria do desempenho de modelos de Aprendizado de Máquina aplicados à estimativa do estoque de carbono orgânico nos solos na região da Amazônia Central.

Para investigar essa questão, este estudo realiza uma análise comparativa de três técnicas de otimização de hiperparâmetros — *Grid Search* (GS), *Random Search* (RS) e *Successive Halving Search* (SH) — aplicadas a três diferentes modelos de AM — Árvore de Decisão (AD), *Random Forest* (RF) e *Gradient Boosting* (GBoost). Os resultados obtidos são avaliados em relação às métricas de desempenho Raiz do Erro Quadrático Médio (RMSE) e Média dos Erros Absolutos (MAE) e comparados com os resultados sem a utilização da estratégia de otimização.

Além disso, este trabalho preenche uma lacuna na literatura, já que até o momento não foram identificados estudos que combinem uma abordagem centrada em dados e a otimização de hiperparâmetros no contexto da determinação do estoque de carbono em solos brasileiros. A expectativa é que as análises realizadas demonstrem como a aplicação de estratégias de otimização de hiperparâmetros pode melhorar significativamente o desempenho dos modelos, contribuindo para avanços metodológicos na estimativa de SOCS.

O presente trabalho está organizado em mais 5 capítulos. No Capítulo II é apresentado o referencial teórico que forma a base para o bom entendimento do trabalho desenvolvido. O Capítulo III apresenta os trabalhos relacionados. O Capítulo IV explica a metodologia e a base de dados utilizada. Os resultados são discutidos no Capítulo V e, por fim, o Capítulo VI realiza as considerações finais.

Capítulo II Referencial Teórico

O objetivo deste capítulo é abordar os conceitos que fundamentam esta pesquisa. Para tanto, ele está dividido em oito seções: a Seção II.1 relata a importância do ciclo do carbono; a Seção II.2 discute sobre os mapeamentos digitais do solo tradicional e digital; a Seção II.3 revisa algumas técnicas de AM que podem ser aplicadas na solução do problema em estudo; a Seção II.4 apresenta uma breve descrição sobre a técnica de validação cruzada; a Seção II.5 aborda sobre a seleção de atributos, avaliando quais destes são os mais relevantes para o modelo; a Seção II.6 apresenta de forma sucinta as técnicas de otimização de hiperparâmetros; e por fim, a Seção II.7 discute as métricas usadas para avaliar os modelos.

II.1 Ciclo global do carbono

Em ecossistemas terrestres, a matéria orgânica do solo (MOS) é importante reservatório de nutrientes, energia, e principalmente, de carbono orgânico. Sem a presença da matéria orgânica, a superficie terrestre seria uma mistura estéril de minerais em processo de desagregação e decomposição. A MOS é um componente importante do solo, afetando diversos processos físicos, químicos e biológicos, desempenhando importantes funções [Leite, 2004].

O carbono orgânico no solo é um componente essencial dos ecossistemas terrestres, representando o carbono presente em compostos orgânicos no solo, principalmente na forma de matéria orgânica decomposta. O MOS reflete o equilíbrio dinâmico entre o carbono adicionado ao solo pela vegetação e aquele perdido para as águas profundas ou liberado na atmosfera [Romão, 2012; Teixeira et al., 2017].

Esse componente é vital para a saúde do solo e desempenha um papel central em diversos processos ambientais e agrícolas. Entre suas funções destacam-se: o fornecimento de nutrientes essenciais às plantas, como nitrogênio e fósforo; a melhoria da capacidade do solo em reter água; a formação de agregados que conferem estrutura e estabilidade ao solo; e, de particular relevância para este estudo, o papel do MOS como um importante reservatório de carbono. Essa última função contribui para a mitigação das mudanças climáticas, ao reduzir as emissões de CO_2 na atmosfera, que são intensificadas por práticas como desmatamento, queimadas e agricultura intensiva sem manejo adequado, como a ausência de rotação de culturas. O aumento dessas emissões intensifica

o efeito estufa, impactando negativamente a segurança alimentar global [Boschiero, 2020; Crespo et al., 2023].

O ciclo global do carbono, conforme ilustrado na Figura II.1, descreve o movimento do carbono entre diferentes reservatórios da Terra, incluindo a atmosfera, os oceanos, a biosfera e a litosfera. Esse processo envolve diversas etapas interligadas, como a fotossíntese, a respiração, a decomposição, a erosão e a sedimentação [Grace, 2004; Ussiri and Lal, 2017]. Essas interações complexas ressaltam a importância do MOS não apenas para a saúde do solo, mas também para o equilíbrio climático e ecológico do planeta.

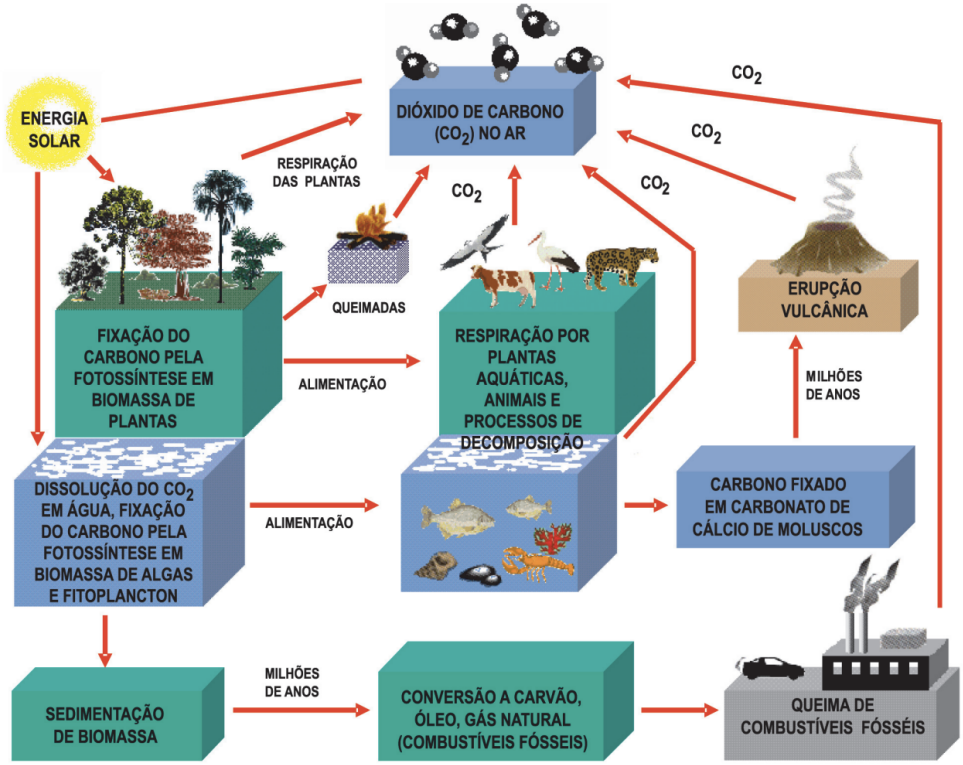


Figura II.1: Ciclo global do carbono. Fonte: [Martins et al., 2003].

Assim, a obtenção de informações detalhadas e precisas sobre a distribuição espacial do carbono no solo é essencial. Nesse contexto, diferentes abordagens têm sido empregadas para mapear e caracterizar os solos, incluindo métodos tradicionais e técnicas mais avançadas, como o mapeamento digital do solo. Essas estratégias oferecem ferramentas importantes para compreender a variabilidade espacial do MOS e auxiliar na tomada de decisões para a gestão sustentável do solo e do carbono. A seguir, serão discutidos os principais aspectos e avanços dessas abordagens.

II.2 Mapeamento de solos: tradicional e digital

Os levantamentos de dados de solos apresentam não apenas a distribuição espacial das diferentes classes de solos, mas também servem para uma descrição e interpretação, fornecendo informações

importantes. A dificuldade do levantamento de solos consiste em mapear algo que não se vê [Hofig, 2014]. Segundo Guerri et al. [2024], os métodos tradicionais que envolvem medições de campo, investigações e análises estatísticas são trabalhosos, demorados e caros.

Os métodos de levantamento de solos tradicionais são demorados e não apresentam adequadamente as informações demandadas pelos usuários. Há problemas de precisão da informação e confiabilidade das interpretações qualitativas. Por esses motivos, o mapeamento tradicional do solo vem sendo criticado ao longo do tempo. Mesmo assim, os levantamentos tradicionais são ainda a principal fonte de informações sobre os solos no Brasil [Lima et al., 2013].

Atualmente, a aplicação de técnicas pedométricas tem sido amplamente utilizada para descrever a variabilidade espacial do solo, consolidando-se por meio do Mapeamento Digital do Solo (MDS). Essas técnicas combinam conhecimentos de geostatística, ciência do solo e ciência da computação para gerar mapas digitais de solos, fornecendo informações detalhadas sobre a distribuição espacial de diferentes atributos do solo, como textura, pH, matéria orgânica, dentre outras. Esse avanço tem sido impulsionado pela integração com Sistemas de Informação Geográfica (SIG), dados digitais e tecnologias de sensoriamento remoto. A Figura II.2 ilustra como a pedometria se fundamenta na premissa de que a complexidade e as lacunas de conhecimento sobre as propriedades do solo podem ser abordadas por meio da aplicação de métodos matemáticos, numéricos e estatísticos. Essa abordagem destaca-se por sua natureza interdisciplinar, reunindo conhecimentos de diversas áreas para enfrentar os desafios associados ao estudo do solo [Lima et al., 2013].



Figura II.2: A interdisciplinaridade da Pedrometria. Fonte: [Lima et al., 2013].

No entanto, a crescente complexidade dos dados espaciais e a necessidade de modelagens mais robustas têm levado à adoção de abordagens mais sofisticadas, como o uso de algoritmos de Aprendizado de Máquina (AM). Essas técnicas oferecem novas possibilidades para lidar com a variabilidade e a incerteza inerentes às propriedades do solo, representando um passo à frente na busca por

soluções mais eficientes e escaláveis. A seguir, serão explorados os principais conceitos e aplicações do Aprendizado de Máquina no contexto do MDS e sua contribuição para o avanço dessa área.

II.3 Aprendizado de Máquina

O Aprendizado de Máquina (AM) é uma área de pesquisa dentro da Inteligência Artificial (IA) que utiliza técnicas computacionais para modelar dados e resolver problemas complexos. As técnicas de AM consistem em uma classe de algoritmos baseados em dados, projetados para mineração de dados e reconhecimento de padrões. Esses algoritmos são amplamente aplicados em diversos campos da ciência, sendo particularmente eficazes na solução de problemas de regressão e classificação. Esses problemas envolvem a análise de grandes volumes de dados para identificar padrões e realizar predições com base nesses padrões, demonstrando a versatilidade e o potencial do AM em diferentes contextos [Mitchell, 1997; Faceli et al., 2011; Wadoux et al., 2020a].

De acordo com Aqdam et al. [2022], "o AM é um método de análise de dados que automatiza uma estrutura de modelo analítico. É uma subárea da IA baseada na ideia de que os sistemas podem aprender com os dados, reconhecer padrões e tomar decisões com a menor intervenção humana".

Os problemas em Aprendizado de Máquina podem ser categorizados em supervisionados e não supervisionados, dependendo do tipo de dado e objetivo da análise. No aprendizado supervisionado, amplamente utilizado, o modelo é treinado com um conjunto de dados em que os valores da variável de interesse (variável resposta) já são conhecidos. O objetivo é aprender a relação entre as variáveis preditoras e a variável resposta, permitindo ao modelo realizar predições precisas para novos dados [Izbicki and Santos, 2020]. Exemplos comuns incluem problemas de classificação, como predizer a categoria de um e-mail (*spam* ou não), e regressão, como estimar o preço de uma casa com base em suas características.

Por outro lado, o aprendizado não supervisionado trabalha com dados em que não há rótulos ou valores conhecidos para a variável resposta. Nesse caso, o objetivo é identificar padrões ocultos ou estruturas nos dados. Técnicas não supervisionadas são frequentemente usadas em agrupamentos (*clustering*), como segmentar clientes com base em comportamentos semelhantes, e na redução de dimensionalidade, como o uso da Análise de Componentes Principais (PCA) para simplificar conjuntos de dados complexos [IBM, 2024].

Enquanto o aprendizado supervisionado é adequado para tarefas em que se conhece o resultado esperado, o aprendizado não supervisionado é ideal para explorar dados sem rótulos e ter percepções que não são imediatamente aparentes. Ambas categorias desempenham papéis complementares na análise e modelagem de dados em AM.

Uma das primeiras etapas na análise de um problema de Aprendizado de Máquina (AM) é

determinar se ele envolve regressão ou classificação, pois essa distinção orienta a escolha das técnicas e algoritmos mais adequados. Os problemas de regressão surgem quando o objetivo é predizer um valor numérico contínuo, como é o caso do SOCS abordado neste trabalho. Por outro lado, os problemas de classificação envolvem a predição de categorias, ou seja, a atribuição de uma observação a uma classe dentro de um conjunto limitado de opções. Embora possam incluir valores numéricos, nesses casos, esses valores representam variáveis categóricas, correspondendo a diferentes classes [Science, 2024].

Para entender melhor as tarefas de AM, a Figura II.3 apresenta uma hierarquia que ilustra a estrutura dessas abordagens. No topo está o aprendizado indutivo, que consiste no processo de generalização a partir dos dados. No segundo nível, encontram-se os dois principais tipos de aprendizado: o supervisionado (também conhecido como preditivo), voltado para tarefas como regressão e classificação, e o não supervisionado (ou descritivo), que explora padrões e estruturas ocultas nos dados [Faceli et al., 2011]. Essa organização permite compreender como diferentes técnicas de AM podem ser aplicadas a uma ampla gama de problemas, garantindo escolhas mais assertivas e alinhadas aos objetivos do estudo.

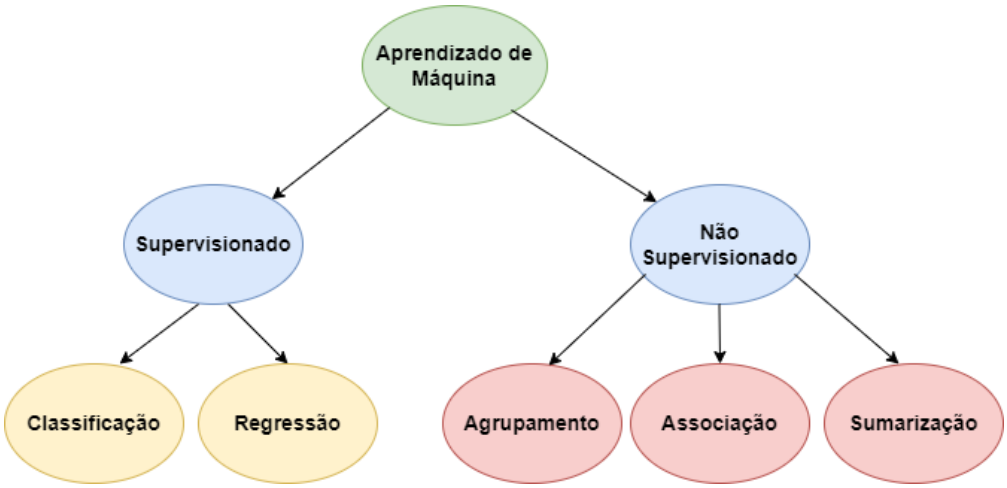


Figura II.3: Hierarquia do Aprendizado de Máquina. Adaptado de: [Faceli et al., 2011].

O desenvolvimento de um modelo de AM segue alguns passos, em geral, descritos pelas etapas de coleta e preparação dos dados, seleção de variáveis/atributos, escolha do algoritmo, determinação dos parâmetros/hiperparâmetros do algoritmo, treinamento e avaliação do desempenho do modelo construído. Seguir estas etapas auxilia no processo de encontrar um modelo de AM de melhor desempenho preditivo.

Apesar de fornecer maior acurácia de predição do que outros modelos estatísticos convencionais, os modelos de AM são considerados uma caixa preta. Contabilizar o conhecimento especializado existente sobre o solo com AM é um exercício desafiador e é resolvido integrando o conhecimento especializado. Por isso é importante a consulta de um especialista ao construir um modelo de

AM [Wadoux et al., 2023].

O foco deste trabalho encontra-se no aprendizado supervisionado envolvendo regressão. Neste tipo de abordagem, o objetivo é aprender uma função que mapeia entradas para saídas com base nos exemplos fornecidos durante o treinamento. Dentre os algoritmos desta classe, este trabalho utiliza-se dos algoritmos de *(i)* Árvore de Decisão (AD), *(ii) Random Forest* (RF), e *(iii) Gradient Boosting* (GBoost).

II.3.1 Árvores de decisão

Uma AD para regressão é um modelo de AM que prediz valores contínuos ao invés de classes discretas. Este modelo constrói uma estrutura hierárquica, na qual as decisões são feitas em cada nó da árvore, dividindo o espaço dos dados em sub-regiões menores. Cada sub-região é associada a um valor preditivo, que geralmente é a média dos valores dos dados de treinamento que caem naquela região. A AD para regressão é uma ferramenta interessante, principalmente pela sua interpretabilidade. No entanto, ela pode ser propensa ao *overfitting* se não for bem ajustada, especialmente em conjuntos de dados com muito ruído. Técnicas de poda e de regularização podem ajudar a melhorar o desempenho dos modelos de AD.

O Algoritmo 1 apresenta um pseudocódigo para um algoritmo de árvore de decisão. Na versão apresentada, o algoritmo chama *BuildTree* recursivamente para construir os ramos esquerdo e direito da árvore. O processo continua até que uma condição de parada seja atendida (como profundidade máxima ou dados homogêneos).

Algorithm 1 Algoritmo de Árvore de Decisão - Construção Recursiva

```
function BuildTree(data, depth)
    Input: data - conjunto de dados de treinamento
           depth - profundidade atual da árvore
    Output: Nó da árvore

    if Condição de parada é atendida then
        return CreateLeaf(data)                    ▷ Retorna uma folha com a classe majoritária
    end if

    bestFeature, bestThreshold ← FindBestSplit(data)          ▷ Escolhe a melhor divisão

    leftData, rightData ← SplitData(data, bestFeature, bestThreshold)

    leftChild ← BuildTree(leftData, depth + 1)  ▷ Constrói o ramo esquerdo recursivamente
    rightChild ← BuildTree(rightData, depth + 1) ▷ Constrói o ramo direito recursivamente

    return CreateNode(bestFeature, bestThreshold, leftChild, rightChild)   ▷ Retorna um nó interno
end function
```

A função *FindBestSplit* (data) é responsável por determinar o melhor atributo e o melhor limiar para dividir os dados. Já a função *SplitData* (data, feature, threshold) realiza a divisão dos dados com base no atributo e limiar selecionados. Criar uma folha contendo a classe majoritária dos dados é a responsabilidade da função *CreateLeaf* (data). Por fim, *CreateNode* (feature, threshold, leftChild, rightChild) cria um nó interno com a melhor divisão e subárvores.

II.3.2 Floresta Aleatória

Uma RF é um algoritmo desenvolvido como uma extensão de árvores de regressão para melhorar a predição da previsão do modelo. No procedimento de treinamento, o algoritmo de RF produz um grande conjunto (*ensemble*) de dados como árvores interdependentes e calcula a média deles [Silatsa et al., 2020; Aqdam et al., 2022].

O algoritmo RF é baseado em AD, onde cada AD é treinada em uma amostra aleatória do conjunto de dados através da técnica *booststrap* (amostragem com reposição) para realizar regressões [Kubrusly, 2024]. Os métodos de amostragem *booststrap* reduzem a sensibilidade de RF ao *overfitting*. Um cenário de *overfitting* ocorre quando, nos dados de treino, o modelo tem um desempenho excelente, porém quando utilizamos os dados de teste o resultado é ruim [Xie et al., 2022].

O Algoritmo 2 foi extraído da *homepage* de Matthew N. Bernstein[1] e apresenta o pseudocódigo do algoritmo *Random Forest*.

II.3.3 Gradient Boosting

GBoost é uma técnica que cria várias árvores de decisão sequencialmente geradas a partir dos erros de predição do modelo da árvore anterior. Essas iterações acontecem por um determinado número de vezes até que o erro seja o menor possível. Ele se assemelha ao algoritmo RF, combinando modelos fracos para criar um modelo mais robusto. A diferença para o RF é que o resultado final precisa passar por cada uma das árvores, na ordem em que elas foram construídas, para ser obtido o modelo final. Essa técnica é conhecida por sua capacidade de melhorar gradualmente a acurácia do modelo à medida que novos modelos são adicionados [Mahmoudzadeh et al., 2020; Kubrusly, 2024].

O Algoritmo 3 apresenta um pseudocódigo do GBoost, conforme descrito por Xinwen Xu [2]:

[1] `https://pages.cs.wisc.edu/~matthewb/pages/notes/pdf/ensembles/RandomForests.pdf`

[2] `https://www.researchgate.net/figure/Pseudo-code-of-gradient-boosting-method-drawn-by-the-author_fig2_368497917`

Algorithm 2 Floresta Aleatória - Construção e Predição

function RANDOMFOREST($data, nTrees, maxFeatures, maxDepth$)
  Input:
    $data$ - conjunto de dados de treinamento
    $nTrees$ - número de árvores na floresta
    $maxFeatures$ - número máximo de características usadas em cada nó
    $maxDepth$ - profundidade máxima das árvores
  Output: Floresta de árvores

  $forest \leftarrow []$ ▷ Inicializa a floresta vazia
  for $i = 1$ to $nTrees$ **do**
    $bootstrapSample \leftarrow$ **BootstrapSample**($data$) ▷ Gera amostra com reposição
    $tree \leftarrow$ **BuildTree**($bootstrapSample, maxFeatures, maxDepth$) ▷ Cria uma árvore de decisão
    Append $tree$ to $forest$ ▷ Adiciona a árvore à floresta
  end for
  return $forest$
end function

function PREDICT($forest, instance$)
  Input:
    $forest$ - floresta de árvores
    $instance$ - instância para predição
  Output: Predição da classe ou valor

  $predictions \leftarrow []$
  for $tree$ in $forest$ **do**
    $prediction \leftarrow$ **PredictTree**($tree, instance$) ▷ Obtém predição de cada árvore
    Append $prediction$ to $predictions$
  end for
  return AggregatePredictions($predictions$) ▷ Retorna a moda ou média das predições
end function

Algorithm 3 Descrição do Algoritmo do Gradient Boosting

Input: x ▷ Variável aleatória de entrada
$F^*(x) = \arg\min \sum_{i=1}^{n} \varphi(y_i, \beta)$ ▷ Minimiza a função de perda
for $m = 1$ to $m = M$ **do**
  $\tilde{y}(im) = -\left(\frac{\partial \varphi(yi, F_m(x_i))}{\partial F(x_i)}\right)$, onde $i \in [1, n]$ e $i \in \mathbb{Z}$ ▷ Pseudo-resíduos
  $\alpha_{(m)} = \arg\min_{\alpha} \rho \sum_{i=1}^{n} (\tilde{y}(im) - \rho h(xi; \alpha))^2$ ▷ Ajuste pelos mínimos quadrados
  $\beta_{(m)} = \arg\min_{\beta} \sum_{i=1}^{n} \varphi(y_i) - F_{(m-1)}(x_i) + \beta h(x_i; \alpha_{(m)})$ ▷ Valor ótimo do coeficiente
  $F_{(m)}(x) = F_{(m-1)}(x) + \left(\beta_{(m)} h(x_i; \alpha_{(m)})\right)$ ▷ A aproximação atual $F_{(m-1)}(x)$ é atualizada separadamente em cada região
end for

II.4 Validação Cruzada *K-fold*

Em AM, você pode usar o método de validação cruzada *K-fold* onde o objetivo é estimar o desempenho de um modelo em um conjunto de dados independente, em vez de apenas avaliar o modelo no conjunto de dados de treinamento. Os dados de entrada são dividido em subconjuntos de dados k (também chamados de *folds*). Um modelo é treinado em todos os subconjuntos, menos em um $(k-1)$ dos conjuntos de dados e, em seguida, o modelo que não foi usado para treinamento é avaliado. Esse processo é repetido k vezes.

A Figura II.4 mostra um exemplo de subconjuntos de treinamento e subconjuntos de teste complementar gerados para cada um dos cinco modelos que são criados e treinados durante uma validação cruzada *5-fold*. O modelo um usa os primeiros 20% dos dados para teste e os 80% restantes para treinamento. O modelo dois usa o segundo subconjunto de 20% (20% a 40%) para teste, e os quatro subconjuntos restantes de dados para treinamento e assim por diante, até chegar ao quinto subconjunto de 20% (80% a 100%). Os resultados de cada *fold* (dobra) são então calculados para produzir uma estimativa geral de desempenho.

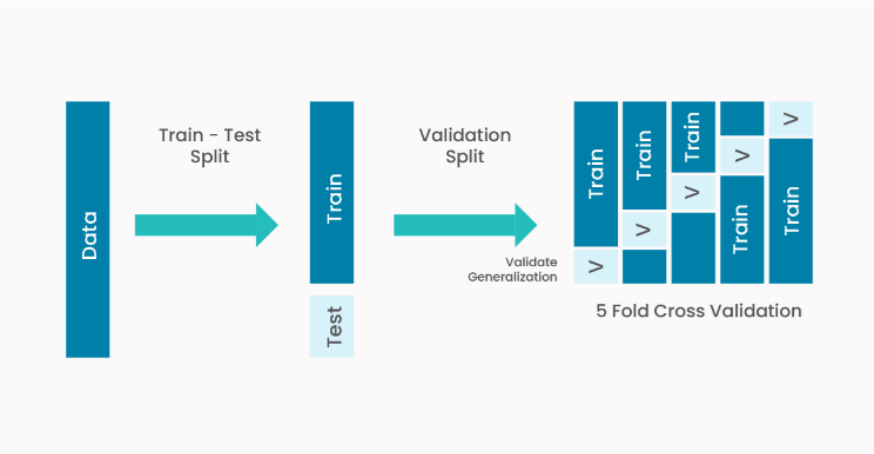


Figura II.4: Validação Cruzada *5-fold*. Fonte: Depuru [2023].

A técnica de validação cruzada (*K-fold*) é utilizada para garantir que o modelo generalize bem para dados desconhecidos, evitando o sobreajuste (*overfitting*) [Anguita et al., 2012].

II.5 Seleção de Atributos

Ao trabalhar com um conjunto de dados, é possível extrair diversos atributos que representam propriedades ou características específicas das observações analisadas. No contexto da ciência de dados, esses atributos, também chamados de *features*, desempenham um papel fundamental na construção de modelos de Aprendizado de Máquina (AM).

A **seleção de atributos** é um processo crucial que tem como objetivo reduzir o número de

variáveis utilizadas no modelo. Essa redução oferece diversos benefícios, como:

- Acelerar o treinamento do modelo de AM;
- Diminuir a complexidade computacional;
- Melhorar a acurácia das predições;
- Mitigar o risco de ajuste excessivo (*overfitting*), ou seja, evitar que o modelo tenha um desempenho inadequado em dados novos e não vistos [Wadoux et al., 2023].

Os métodos de seleção de atributos são comumente classificados em duas principais abordagens, com base nas estratégias utilizadas para identificar os atributos mais relevantes:

1. **Correlação de Pearson**: Mede a relação linear entre os atributos e a variável-alvo.
2. **Abordagem *Wrapper***: Avalia o desempenho do modelo em diferentes subconjuntos de atributos para selecionar os mais relevantes.

Como essas abordagens são complementares, elas podem inclusive ser combinadas para garantir um modelo mais eficiente e robusto.

II.5.1 Correlação de Pearson

A correlação é uma medida estatística que avalia a direção e a intensidade da relação linear entre duas variáveis quantitativas [David S. et al., 2007]. Uma das técnicas amplamente utilizadas para essa finalidade e com baixo custo computacional é a **Correlação de Pearson**, que calcula o grau de associação linear entre duas variáveis com base em sua covariância.

O coeficiente de Correlação de Pearson (r) varia de -1 a 1, indicando a força e a direção da relação:

- $\mathbf{r = 1}$: Indica uma correlação linear perfeita e positiva, ou seja, à medida que uma variável aumenta, a outra também aumenta proporcionalmente.
- $\mathbf{r = -1}$: Indica uma correlação linear perfeita e negativa, ou seja, à medida que uma variável aumenta, a outra diminui proporcionalmente.
- $\mathbf{r = 0}$: Indica ausência de correlação linear entre as variáveis, embora possam existir outros tipos de relações não lineares.

Valores de r mais próximos de 1 ou -1 indicam uma relação linear forte entre as variáveis, enquanto valores mais próximos de 0 sugerem uma relação linear fraca ou inexistente. É importante destacar que o coeficiente de Pearson mede exclusivamente relações lineares, não sendo apropriado para descrever associações não lineares entre variáveis.

II.5.2 Métodos *Wrapper*

Apesar do método *Wrapper* ter um alto custo computacional, especialmente a pesquisa exaustiva, a vantagem da escolha desse método é por ser adaptado ao modelo de AM para avaliar a seleção de atributos [Daud et al., 2025]. Outra vantagem é que emprega algoritmos evolutivos, enraizado na teoria da otimização, permitindo um processo de pesquisa iterativo que equilibra exploração e aproveitamento [Shahidian et al., 2024]. Além de reduzir a quantidade de atributos visando melhorar a acurácia e também reduzir o sobreajuste (*overfitting*) e a complexidade computacional, o método *Wrapper* aborda dois temas entre os principais desse trabalho, AM e otimização, o que justifica a escolha para avaliar a seleção de atributos.

O método *Wrapper* depende da inferência feita por um modelo de AM calibrado. Ao recalibrar um modelo de AM várias vezes, cada vez removendo o atributo menos importante, pode-se esperar reduzir o número total de atributos com pouca ou nenhuma diminuição na acurácia da predição do modelo [Wadoux et al., 2020a].

Esse método envolve a seleção de atributos em torno do algoritmo de AM e usa a acurácia do desempenho como critério de avaliação. Ele seleciona os atributos mais ideais para o algoritmo de predição, e consequentemente, obtém melhor desempenho e alta acurácia [Zebari et al., 2020].

Enquanto o algoritmo EFS (*Exhaustive Feature Selection*) realiza uma pesquisa exaustiva em todos os subconjuntos possíveis para encontrar o melhor conjunto de atributos, o algoritmo SFS (*Sequential Feature Selection*) que consiste em um método sequencial, começa com uma única variável e continua adicionando mais variáveis, objetivando obter o melhor desempenho do modelo [Galli, 2023].

II.6 Otimização de Hiperparâmetros

A otimização de hiperparâmetros é um passo importante no desenvolvimento de modelos de AM, auxiliando no desenvolvimento de modelos mais precisos e eficientes. Os hiperparâmetros são parâmetros cujos valores são definidos antes do início do processo de treinamento do modelo e não são ajustados durante a aprendizagem. Eles incluem, por exemplo, a taxa de aprendizagem, o número de neurônios em uma camada de uma Rede Neural (RN), entre outros que serão citados nesta Seção [Jamieson and Talwalkar, 2016; Kumar et al., 2018; Agrawal, 2021; Japa et al., 2023; Soper, 2023; Pinna et al., 2023].

No contexto da otimização de hiperparâmetros para essa aplicação, o objetivo é encontrar a combinação ideal de hiperparâmetros que melhore a acurácia do modelo preditivo. Os profissionais geralmente ajustam esses hiperparâmetros usando métodos padrão de força bruta, como a pesquisa em uma grade de hiperparâmetros (*grid search*) ou a amostragem aleatória de hiperparâmetros

(*random search*). Entre os métodos existentes para acelerar a otimização de hiperparâmetros, a redução sucessiva pela metade (*successive halving search*) surgiu como um algoritmo de parada antecipada popularmente utilizada [Li et al., 2020].

A Figura II.5 apresenta os três métodos de otimização de hiperparâmetro supracitados, cada ponto representa uma configuração de hiperparâmetro específica e as cores mais quentes (vermelho) correspondem a um melhor desempenho. Na prática, a aplicação de soluções de aprendizado de máquina requer um ajuste cuidadoso dos hiperparâmetros pertencentes ao modelo para obter alta acurácia preditiva [Li, 2018].



Figura II.5: Otimização por hiperparâmetro. Fonte: [Li, 2018].

O *Grid Search* (GS) envolve a definição de um espaço de busca (*grid*) composto por um conjunto de valores possíveis para cada hiperparâmetro e a avaliação exaustiva de todas as combinações possíveis desses valores, o que pode tornar tal processo computacionalmente custoso [Song et al., 2022]. O *Random Search* (RS) seleciona aleatoriamente combinações de valores de hiperparâmetros a partir de um espaço de busca predefinido, oferecendo uma solução mais eficiente em termos de tempo computacional [Japa et al., 2023].

Já a *Successive Halving Search* (SH) é um método de otimização de hiperparâmetros iterativo baseada em parada antecipada no qual o número de configurações diminui de uma iteração para a próxima, onde o algoritmo usa um pequeno número de casos de treinamento para rapidamente identificar e eliminar configurações com baixo desempenho. As configurações com desempenho inferior (geralmente a metade) são descartadas, mantendo-se apenas as melhores. Ele repete o processo com as configurações mais promissoras até que se obtenha a melhor configuração [Luthra and Janghu, 2023; Soper, 2023].

Como todas essas técnicas impactam o modelo desenvolvido é importante avaliar os resultados obtidos por meio de métricas específicas, no presente trabalho utilizaremos o Erro Médio Absoluto (MAE) e a Raiz do Erro Quadrático Médio (RMSE), que são brevemente apresentadas a seguir.

II.7 Avaliação da Acurácia dos Modelos

Nesta seção, serão apresentadas as principais métricas utilizadas para avaliar o desempenho de modelos de regressão em algoritmos de Aprendizado de Máquina (AM). Essas métricas são essenciais para analisar a qualidade das predição e identificar a eficácia dos modelos desenvolvidos.

Duas das medidas mais amplamente empregadas para avaliar a acurácia de modelos de regressão são o Erro Médio Absoluto (MAE) e a Raiz do Erro Quadrático Médio (RMSE), cujas fórmulas estão descritas nas Eqs. II.1 e II.2, respectivamente. Ambas as métricas fornecem uma síntese da diferença média entre os valores observados e preditos, sendo consideradas entre as melhores medidas gerais de desempenho do modelo devido à sua clareza e abrangência [Ferreira et al., 2023].

$$MAE = \frac{1}{n}\sum_{i=1}^{n} \mid y_i - \hat{y}_i \mid^2 \tag{II.1}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{II.2}$$

Onde n é o número de observações, y_i e $\hat{y}_i$ são os valores observados e preditos, respectivamente, e $\overline{y}$ é a média dos valores observados.

O MAE mede a média das diferenças absolutas entre os valores preditos e observados, refletindo diretamente o erro médio sem considerar sua direção. Já o RMSE incorpora o quadrado das diferenças, penalizando erros maiores de forma mais intensa e, assim, destacando discrepâncias significativas nos dados. Ambas as métricas fornecem uma visão complementar sobre o desempenho dos modelos, auxiliando na escolha e ajuste das soluções mais adequadas para os problemas de regressão.

Por fim, é importante compreender que aplicar a otimização de hiperparâmetros na determinação do SOCS envolve o uso de técnicas de validação cruzada para assegurar que os modelos não apenas se ajustem bem aos dados de treinamento, mas também generalizem bem para dados não vistos. Isso pode garantir a robustez e a acurácia das estimativas de carbono no solo, as quais podem auxiliar na formulação de políticas ambientais e práticas de manejo sustentável.

Capítulo III Trabalhos Relacionados

Este capítulo apresenta os trabalhos encontrados na literatura a partir de um mapeamento sistemático sobre determinação de estoque de carbono orgânico em solos usando técnicas de Aprendizado de Máquina.

III.1 Base de dados e String de Busca

Para identificar o grupo primário de artigos, uma pesquisa foi realizada na base *Scopus* utilizando a *string* de busca: TITLE-ABS-KEY ("*soil*" and "*machine learning*" and "*organic carbon*"), onde foram encontrados 829 artigos. Visando limitar apenas os artigos mais recentes, foram considerados no termo de busca os anos de 2020 a 2024, resultando em 714 artigos base. Do total desses 714 artigos pesquisados, 121 foram selecionados pois estavam dentro do escopo desse trabalho. Utilizando a técnica de *snowballing* foram encontrados mais 5 artigos aplicados em solos.

A Figura III.2 ressalta a recente importância do tema desse trabalho onde apresenta um aumento de trabalhos relacionados nos últimos anos.

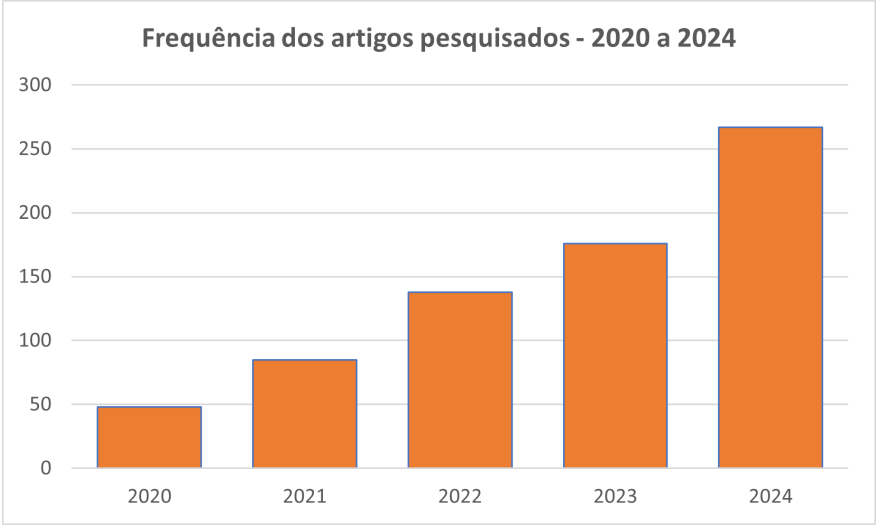
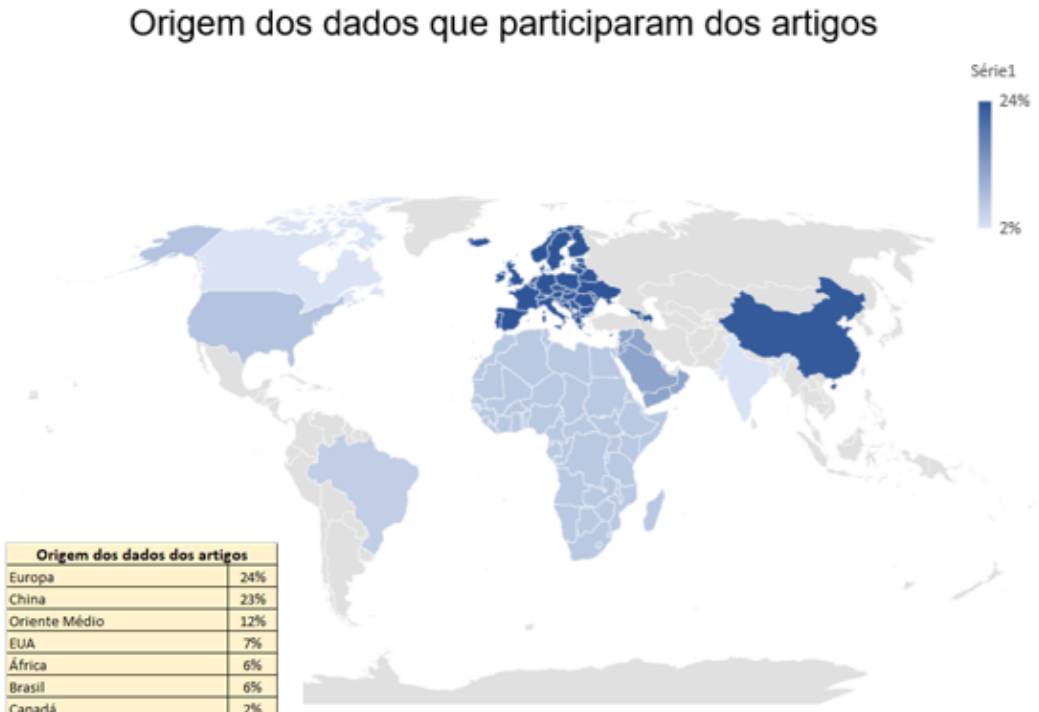


Figura III.1: Frequência dos artigos pesquisados - 2020 a 2024.. Fonte: Scopus - pesquisa em 10/01/2025.

III.2 Origem dos dados dos solos

A Figura III.2 mostra de onde foram extraídas as informações sobre o solo para realizar os estudos referentes aos artigos pesquisados. A maior parte dos dados do solo veio da Europa (24%), seguido da China (23%) e Oriente Médio (12%). As técnicas de AM no contexto das ciências do solo sobre estoque de carbono são usadas em muitos países ao redor do mundo, mas estão concentradas principalmente nos países desenvolvidos como os da Europa, onde fica evidente a relação entre ciência, tecnologia e desenvolvimento. Nos artigos pesquisados, 6% foram em solos brasileiros.



Origem dos dados dos artigos	
Europa	24%
China	23%
Oriente Médio	12%
EUA	7%
África	6%
Brasil	6%
Canadá	2%
Índia	2%
Outros / sem dados	17%
TOTAL	**100%**

Figura III.2: Origem dos dados dos solos pesquisados. Fonte: O próprio autor.

III.3 Métodos de AM mais comumente usados nas ciências do solo

Da enorme variedade de modelos de AM disponíveis, encontramos aproximadamente 20 variantes diferentes que foram aplicadas nas ciências do solo. A maioria dos artigo usou mais de um modelo de AM para fins de comparação. Também houve comparações de modelo de AM com modelos de regressão usados na Estatística, como por exemplo, Modelo Linear Generalizado (GLM), Regressão Linear Múltipla (RLM) e Regressão Bayesiana (EBKRP).

A Tabela III.1 mostra que 89 trabalhos usaram o algoritmo RF, que representa 70,6% dos artigos pesquisados. Seguido da Máquina de Vetor de Suporte (SVM) com 34,1%, Redes Neurais (RN) com 23,8%, e Regressão Cubista (RC) e GBoost, ambos com 19,0%. E depois, AD com 7,9%

dos artigos selecionados.

Tabela III.1: Métodos de Aprendizado de Máquina mais comumente usados.

Frequência e Proporção dos modelos de AM		
RF	89	70,6%
SVM	43	34,1%
RN	30	23,8%
Cubista	24	19,0%
GBoost	24	19,0%
AD	10	7,9%

III.4 Métricas de desempenho mais comumente usadas nas ciências do solo

As principais métricas de desempenhos encontradas nos artigos selecionados foram: RMSE, Coeficiente de Determinação (R^2) e MAE, respectivamente.

A Raiz do Erro Quadrático Médio (RMSE) é uma métrica de avaliação amplamente utilizada nos modelos de AM para medir o desempenho de modelos de regressão. A RMSE é sensível a valores extremos (*outliers*), o que pode distorcer a métrica. No entanto, é importante lembrar que um RMSE baixo pode variar e depende do contexto. Um RMSE no valor de 1000 pode ser ótimo ao tentar predizer o preço de um terreno, mas seria terrível em um caso de predição de estoque de carbono orgânico.

O Erro Médio Absoluto (MAE) também é uma métrica utilizada para avaliar o desempenho de modelos de regressão em AM e avalia a média da diferença entre o valor real com o predito. Ao contrário do RMSE, o MAE não é afetado pelos *outliers*. Interpretando, se o MAE for igual 0,8 na predição de estoque de carbono, como por exemplo em Ferreira et al. [2023] para profundidade de 30cm usando RF, significa que o modelo tende a subestimar ou superestimar o valor de 0,8 desse estoque. Essa é uma forma intuitiva de interpretar a MAE, mas não se deve confundir em estabelecer um intervalo de confiança para as predições.

Existe uma grande variedade de métricas de desempenho disponível, e encontramos 26 variantes diferentes que foram aplicadas na área de ciência do solo. A maioria dos artigos usou mais de uma métrica para fins de comparação.

A Tabela III.2 mostra que 101 usaram a RMSE para calcular a acurácia do modelo, que representa 80,2% dos artigos selecionados. Seguido do R^2 com 71,4% e do MAE com 35,7%. E depois, Coeficiente de Correlação de Concordância de Lin (CCCLin) com 23,8%. A métrica CCCLin é uma medida de concordância das predições do modelo e dos valores medidos [Rentschler et al., 2020].

Tabela III.2: Métricas de desempenho mais comumente usadas.

Frequência e Proporção das métricas de desempenho		
RMSE	101	80,2%
R2	90	71,4%
MAE	45	35,7%
CCCLin	30	23,8%

A RMSE e a MAE são comumente usadas em modelos de regressão, seja nos algoritmos de AM ou em modelos estatísticos. Esse uso comum das métricas de desempenho está relacionado às facilidades de calcular, entender e interpretar os resultados. Ambos fornecem uma medida intuitiva, já que os resultados são expressos nas unidades de medida originais dos dados observados. A diferença é que a MAE ajuda a minimizar a influência dos *outliers* no conjunto de dados, que podem estar impactando nas predições do modelo.

Já o uso do R^2 em AM é um problema, pois o R^2 e a regressão não linear não atuam bem juntas. Ele funciona matematicamente apenas para modelos de regressão linear, apesar de alguns trabalhos ainda apresentarem o uso do R^2 em modelos não lineares. Apesar desse problema, a maioria dos *softwares* estatísticos ainda calcula o R^2 para modelos não lineares e essa prática questionável pode causar interpretações equivocadas sobre quais modelos não lineares são melhores [Book and Young, 2006; Aloe et al., 2010; Tellinghuisen and Bolster, 2011; Nakagawa and Schielzeth, 2013].

III.5 Frequência dos artigos referente aos anos encontrados no *Scopus*

Em relação aos 126 artigos selecionados na pesquisa realizada no *Scopus* referentes aos anos de 2020 até 2024, na revisão da literatura realizada, observamos um aumento no número de publicações de 13 artigos em 2021 para 34 artigos em 2022. Depois, os anos de 2023 e 2024 mantiveram, aproximadamente, o quantitativo de artigos selecionados de 2022.

III.6 Regressão ou classificação: o mais comumente usado nas ciências do solo

Verificando os 126 artigos, na Tabela III.3 constatamos que aproximadamente 85 (67,5%) usaram somente modelos de regressão e 4,8% usaram somente classificação. Analisamos também que 9,5% dos trabalhos usaram tanto a regressão quanto a classificação e 5,6% citaram os dois mas não ficou claro qual foi usado no modelo. Por fim, 7,1% dos artigos não citaram nenhum dois, sejam pelos artigos serem somente textual (revisão) ou simplesmente modelaram mas não deixaram claro qual método foi utilizado.

Tabela III.3: Uso de Regressão e/ou Classificação.

Usou Regressão ou Classificação?		
Somente Regressão	85	67,5%
Somente Classificação	6	4,8%
Regressão e Classificação	12	9,5%
Citou os 2 sem definir qual	7	5,6%
Não citou/revisão	16	7,1%

III.7 Modelo de AM que obteve melhor resultado

Entre os 126 artigos selecionados, pela Tabela III.4, 42 (33,3%) concluíram que RF obteve o melhor desempenho. Em seguida, foi o SVM como o melhor desempenho para 7,9% dos artigos. Constatamos que 46 (36,5%) não fizeram comparação, por usarem somente um modelo ou não apresentarem resultados de comparação.

Verifica-se que na Tabela III.4 somam-se 134 artigos, pois o RF, entre outros, dividiram com outros modelos o resultado de melhor desempenho.

Tabela III.4: Algoritmos de AM de melhor desempenho

Frequência e Proporção dos modelos AM de melhor desempenho		
RF	42	33,3%
SVM	10	7,9%
GBoost	7	5,6%
Cubista	6	4,8%
RN	4	3,2%
Outros modelos	19	15,1%
Não houve comparação	46	36,5%

III.8 Número de estimadores usados no modelo *Random Forest*

O número de estimadores é um parâmetro usado para determinar o número de árvores que será usado no modelo RF. Na Tabela III.5 mostra que o número de estimadores mais usado nos modelos de RF foi o 500, com 8,7% dos 126 artigos selecionados. O número de estimadores não foi especificado em 41,3% dos artigos selecionados, que representa 58,4% dos 89 artigos que usaram o modelo RF.

Tabela III.5: Quantidade de estimadores usados no RF.

Frequência e proporção de n_estimadores		
Não Especificado	52	41,3%
Não usou RF	37	29,4%
50	2	1,6%
100	1	0,8%
200	2	1,6%
500	11	8,7%
600	3	2,4%
700	1	0,8%
1000	1	0,8%
10000	1	0,8%
Outros	13	10,3%

III.9 Método de validação cruzada (*K-fold*)

Na Tabela III.6, analisando todos os 126 artigos, 41,3% usaram validação cruzada com valor de $K = 10$ e 9,5% usaram com $K = 5$. Os 13,5% dos trabalhos usaram validação cruzada mas não especificaram o valor do K. Os trabalhos que não usaram validação cruzada foram 33,3% do total.

Tabela III.6: Validação Cruzada utilizada com os valores de K.

Frequência e proporção de *K-fold*		
Não usou	42	33,3%
K não divulgado	17	13,5%
10	52	41,3%
8	1	0,8%
5	12	9,5%
4	1	0,8%
3	1	0,8%

Como exemplo temos o trabalho de [Ferreira et al., 2023]. Para avaliar o modelo de área total em uma área remota da floresta Amazônica, aplicaram validação cruzada *K-fold* ($K = 5$). Dividiram aleatoriamente os dados em cinco partições de tamanho igual. Para cada partição n, treinaram o modelo proposto nas quatro partições restantes e o testaram na partição n. A nota final foi a média das cinco notas obtidas.

III.10 Levantamentos dos dados de solos utilizados

O MDS é um método de levantamento e mapeamento de solos baseado em um modelo solo-paisagem usando análise espacial e métodos matemáticos [Wan et al., 2022]. Todos os artigos, excluindo os de revisões, usaram dados de solo e algum modelo matemático, seja de AM ou modelo Estatístico, independente se o levantamento do solo foi tradicional ou digital. Todos esses artigos usaram o método MDS.

A Tabela III.7 mostra que, entre os 126 artigos selecionados, 54,8% usaram algum método de sensoriamento remoto, como por exemplo, espectroscopia de infravermelho ou mapa de elevação digital. Verifica-se que 18,3% dos artigos usaram algum Sistema de Informação Geográfica (SIG), que é todo *software* ou *hardware* capaz de criar, coletar, processar, editar e armazenar dados dos solos. Como por exemplo de banco de dados usados temos o LUCAS da Eurostat e o SRTM da NASA. Por fim, 19,8% usaram método tradicional para o levantamento de dados do solo, como a remoção do solo *in loco* usando amostragem.

Tabela III.7: Métodos usados no levantamento de dados do solo.

Frequência e proporção - Levantamento dados do solo		
Sensoriamento Remoto	69	54,8%
Tradicional (remoção do solo)	25	19,8%
Sistemas de Informação Geográfica	23	18,3%
Revisão (sem dados)	9	7,1%

Pela Figura III.3, verifica-se que no levantamento tradicional do solo, todos foram por amostragem e 2 tiveram ajuda do *software* ArcGIS. No SIG, o banco de dados mais utilizado foi o *Land use and land cover survey* (LUCAS) da Eurostat. E por fim, no uso de sensoriamento remoto, o mapa de elevação digital (DEM) e o sensor de satétile foram os mais utilizados nos artigos selecionados.

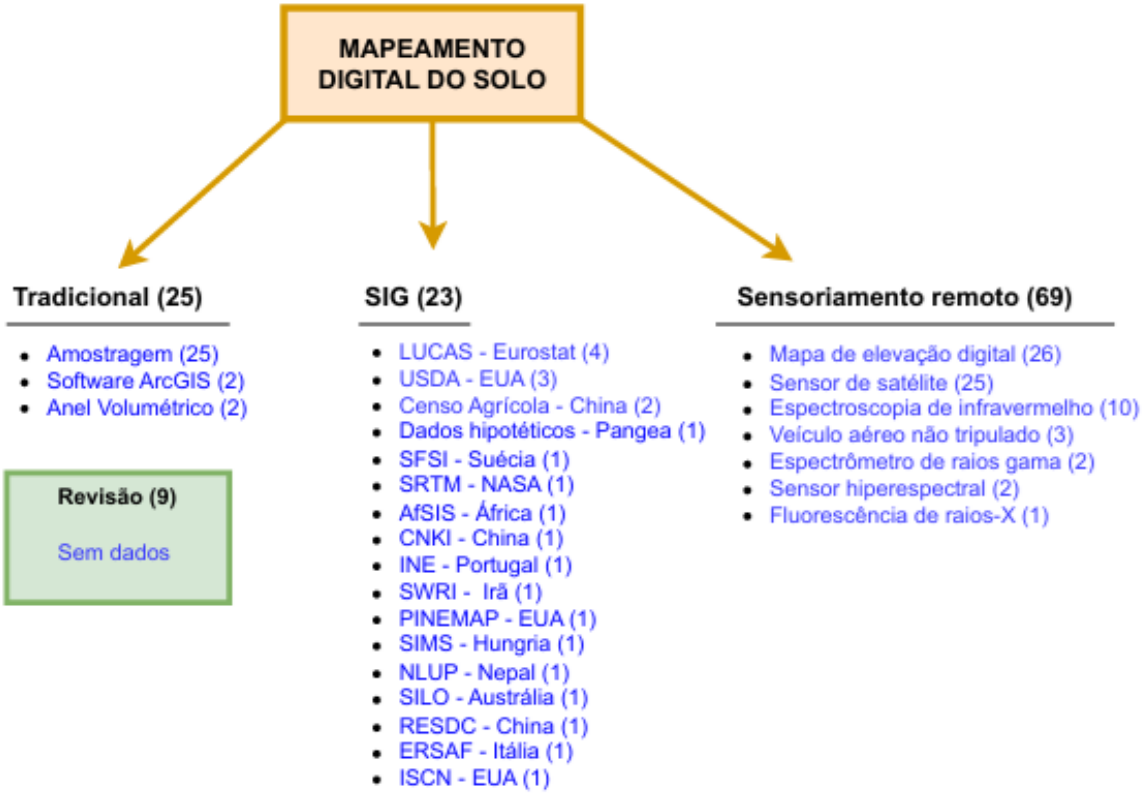


Figura III.3: Levantamento de dados do solo.

III.11 Interpretabilidade dos modelos de AM

A interpretabilidade é a tradução de um modelo em termos compreensíveis por humanos e combina com a plausibilidade do modelo e a descoberta de hipóteses. Os cientistas do solo contam com algoritmos de AM para obter informações sobre os resultados dos modelos utilizados. Apesar de fornecer maior acurácia de predição do que outros modelos convencionais, encontrados na Estatística, os modelos de AM são considerados uma 'caixa preta'. Existe o desafio de comparação do conhecimento dos mapas produzidos por especialistas em solos com os mapas preditos por um modelo de AM, e isso só é possível se o modelo de AM for interpretável por humanos e as relações entre as covariáveis forem realistas (o modelo é plausível) [Wadoux et al., 2023].

Na exaustiva pesquisa dos trabalhos encontrados na literatura, entre os 126 selecionados, somente 4 trabalhos apresentaram uma interpretabilidade sobre os resultados encontrados nos modelos utilizados, 5 trabalhos comentaram sobre a interpretabilidade e somente apresentaram resultados, e o restante, resultados numéricos sem uma análise do comportamento do solo.

Entre os quatro trabalhos que apresentaram uma interpretabilidade do modelo, temos o trabalho de Ross et al. [2020], que teve como objetivo a predição do carbono no solo usando o algoritmo de RF. Ele diz que usou um parcimonioso conjunto de covariáveis para aumentar a interpretação do modelo devido à caixa-preta, que é uma natureza do algoritmo RF. Os autores também abordam o fato de que generalizar corretamente torna-se cada vez mais difícil à medida que o número de covariáveis aumenta. Interpretando através dos resultados de predição usando RF, embora o estoque de carbono varie 0,3 para 8,2 Km m^{-2} na camada superficial do solo e de 0,4 a 17,6 Kg m^{-2} no subsolo, esta variabilidade é previsivelmente distribuída com precipitação, umidade do solo, nitrogênio e conteúdo de areia, emissões de raio gama, temperatura média anual mínima e elevação. Através das predições, os autores estimam que o estoque de carbono orgânico está atualmente estabilizado na floresta de 34,7 milhões de hectares no sudeste dos EUA. Nesse trabalho não houve comparação entre algoritmos pois só foi utilizado o RF.

Outro trabalho que apresentou uma interpretabilidade do modelo foi o apresentado por Ferreira et al. [2023], que teve como objetivo avaliar dois tipos diferentes de abordagem de amostragem para predizer o SOCS nas profundidades de 30 e 100 cm, utilizando três algoritmos de AM, onde RF foi considerado o melhor modelo. Interpretando os resultados, o trabalho mostrou que o uso da seleção prévia de covariáveis, combinada com a abordagem de uma área de referência (RA) situada na Floresta Amazônica, permite desenvolver modelos mais precisos. O trabalho abordou na modelagem de AM três elementos considerados importantes por Wadoux et al. [2020a]: plausibilidade, interpretabilidade e explicabilidade que, para ele, combina a predição do modelo com a explicação pedológica. O trabalho de Ferreira et al. [2023] mostrou que a metodologia que combina

conhecimento pedológico com otimização de parâmetros do algoritmo, seleção prévia de covariáveis e transferibilidade do modelo por meio da abordagem RA foi fundamental para obter resultados satisfatórios, mesmo com a dificuldade de mapeamento da Floresta Amazônica por ser uma área de difícil acesso.

Wadoux et al. [2023] relatam o desenvolvimento de um método estatístico (*Shapley*) para interpretação de modelos complexos, para o estudo da variação de SOCS. Os autores ajustaram um modelo de RF e usaram um conjunto de covariáveis ambientais. Descobriram que a vegetação e a topografia foram, em geral, os fatores mais importantes da variação do SOCS na França. Concluíram que o método *Shapley* constitui uma ferramenta promissora para interpretar modelos e que seu principal valor agregado é permitir uma interpretação dos fatores ambientais que contribuem para uma predição.

Por fim, temos o trabalho de Kebonye et al. [2023] que apresentou uma interpretabilidade do modelo utilizado. Diz que existem vários modelos de caixa preta mais poderosos, como RF sendo utilizados, mas há relativamente poucos estudos que avaliam explicitamente o desempenho de algoritmos de AM, como árvores de decisão, apesar de serem mais inteligentes e mais flexíveis do que o RF. Demonstraram uma maneira simples, mas relevante de melhorar o desempenho de um modelo de árvore de classificação e regressão (CART) para o MDS, ao mesmo tempo que beneficia do seu potencial de inteligibilidade, interpretabilidade e intuição. O trabalho interpretou que os níveis de carbono orgânico do solo foram maiores em áreas florestais de alta altitude na República Tcheca. Explicou também que podemos nos beneficiar da interpretabilidade enquanto melhoramos o desempenho do modelo CART com base em condições simples, como divisão de dados e seleção de recursos.

Entre os trabalhos que citaram a dificuldade da interpretabilidade do modelo e somente apresentaram resultados, temos o de Westhuizen et al. [2023] que comparou o algoritmo RF multivariada (MRF) com RF individual: "Deve ser notado que com o MRF a importância da variável não pode ser interpretada individualmente para cada propriedade do solo no modelo. Isso também foi notado por Wadoux et al. [2020a]". Podemos mencionar também o trabalho de Chen et al. [2021]: "Embora o RF tenha sido o melhor modelo na predição do SOCS, o RF foi limitado em revelar as relações entre o SOCS e as variáveis ambientais, restringindo sua interpretabilidade".

Como exemplos de trabalhos que também apresentaram somente resultados, sem uma interpretação, temos Mahmoudzadeh et al. [2020]; dos Santos et al. [2022]; Yang et al. [2022a] que citaram que os resultados mostraram que o algoritmo RF teve melhor desempenho na predição de SOCS em comparação com os outros algoritmos de AM. Temos os trabalhos de Emadi et al. [2020]; Taghizadeh-Mehrjardi et al. [2020] que citaram que RN superaram os outros algoritmos AM em termos de poder da incerteza de predição. Tivemos também os trabalhos de Xu et al. [2021];

Sakhaee et al. [2022]; Huang et al. [2023] que apresentaram que o algoritmo SVM foi o que obteve melhor resultado se baseando nas métricas de desempenho.

III.12 Os artigos avaliados

A seguir, citaremos todos os artigos selecionados que usaram modelos de AM ou modelos estatísticos, através do mapeamento digital do solo, com as explicações sobre os métodos utilizados no levantamento de dados do solo, os resultados das métricas de desempenho e a interpretabilidade do modelo. Como a RF foi amplamente utilizado nos artigos avaliados, separamos a descrição em duas partes: os que usaram RF, mesmo com outros algoritmos, e os que não usaram RF.

Os 9 artigos Padarian et al. [2020]; Wadoux et al. [2020a]; Odebiri et al. [2021]; Sharma et al. [2022]; Li et al. [2023, 2024a]; Qadri and Ceccato [2024]; Radocaj et al. [2024]; Li et al. [2024b], que fizeram uma revisão abordando os temas sobre ciência do solo, algoritmos de AM, métricas de desempenho e mapeamento digital do solo, não apresentaram um modelo matemático através de um levantamento de dados do solo.

III.12.1 Artigos que usaram RF

Diversos trabalhos vêm sendo desenvolvidos para estimar SOCS e a maioria dos trabalhos utilizou como algoritmo de AM a RF. Para começar, temos os trabalhos de Silatsa et al. [2020]; John et al. [2020]; Chen et al. [2021]; Mousavi et al. [2022]; Hateffard et al. [2023]; Pellikka et al. [2023]; Radocaj et al. [2024]; Wang et al. [2024a]; Ou et al. [2024]; Oliveira et al. [2024], que tiveram como objetivo, predizer o SOCS usando AM e modelos geoestatísticos clássicos como a RLM, Regressão Amplificada Generalizada (GBR) e Regressão dos Mínimos Quadrados Parciais (RMQP). Tiveram outros trabalhos dos autores Junior et al. [2016]; Wanga et al. [2021] que usaram AM e modelo geoestatístico mas as abordagens de mapeamento digital de solo foram relacionadas à matéria orgânica do solo (MOS), que continha carbono orgânico e outros elementos.

Nos trabalhos de Mishra et al. [2020]; Sahu et al. [2021]; Matinfar et al. [2021]; Farooq et al. [2022]; Wan et al. [2022]; Zhu et al. [2022]; Zhang et al. [2023a]; Ou et al. [2024] foram usados AM e Krigagem de Regressão (KR), que é outro método de regressão utilizado em geoestatística para aproximar dados para predizer SOCS. Também para predizer SOCS, em Szatmári et al. [2021] além de AM e KR, foi utilizado também co-simulação estocástica espacial. E o trabalho de Kaya et al. [2022] que apresentou AM e RK para predizer SOCS e o elemento Fósforo.

Alguns autores usaram para predição a função de Pedotransferência (PTF). Além de Junior et al. [2016]; Wanga et al. [2021] já citados anteriormente, temos o trabalho de Mishra et al. [2022a] que teve como objetivo, desenvolver, avaliar e comparar a eficácia dos algoritmos de AM na determinação do capacidade de troca catiônica (CTC) em diferentes usos da terra e de Xiao et al.

[2022], que avaliou o potencial de SOCS associado a minerais usando funções de pedotransferência baseado em AM.

Um dos algoritmos de AM mais utilizados para a predição do SOCS foi a RC, presente nos trabalhos de Meliho et al. [2023]; Radocaj et al. [2024]; Nozari et al. [2024]; Ou et al. [2024]; Moharana et al. [2024]. Podemos também citar como exemplos o trabalho dos autores Rentschler et al. [2020] que mediu as propriedades do solo em 3D fornecidas por medições geofísicas registradas com sensores, onde a predição espacial de SOCS com RC forneceu um erro ligeiramente menor do que com RF, e também, o trabalho dos autores Li et al. [2024c] que apresentou a variação espacial e estimativa de SOCS em terras agrícolas na região do nordeste da China com RF superando o modelo cubista.

Considerando trabalhos que utilizaram modelos de AM para fins de comparação para predição SOCS, onde o levantamento do solo foi por sensoriamento remoto, podemos citar Lieb et al. [2016] em uma paisagem de floresta tropical montanhosa, a partir de preditores derivados de uma imagem de satélite. Em Sothe et al. [2022] os autores testaram uma abordagem de AM tridimensional (3D) para a predição SOCS. O trabalho dos autores Kumar et al. [2023b] preparou um mapa SOCS digital com diferentes algoritmos de ML e avaliou a eficiência dos modelos. Goydaragh et al. [2021]; Xie et al. [2022]; Yang et al. [2022a]; Zhang et al. [2022]; Voort et al. [2023]; Broeg et al. [2023a]; Santos et al. [2023]; Chen et al. [2024]; Golkar Amoli et al. [2024] também avaliaram as métricas de desempenhos de AM para predizer SOCS.

Em Mahmoudzadeh et al. [2020] foram utilizados modelos de AM para fins de comparação para predição SOCS, onde o levantamento do solo foi por dados coletados nos SIG. Os autores utilizaram banco de dados do Instituto de Pesquisa de Solo e Água do Irã. Liu et al. [2022] utilizaram dados do *China National Knowledge Infrastructure* (CNKI), Lamichhane et al. [2022] do repositório do projeto nacional de uso da terra do Governo do Nepal (NLUP), Sakhaee et al. [2022] do inventário de solo agrícola alemão, Fu et al. [2024] do *United States Department of Agriculture* (USDA) dos EUA, Gao et al. [2024] do *Resource and Environmental Science Data Center* (RESDC) da China e Adeniyi et al. [2024] do *Ente Regionale Servizi Agricoltura Foreste* (ERSAF) da Itália.

Em Ladd et al. [2022] é apresentado modelos de AM para predição de SOCS, onde o levantamento do solo foi por dados coletados de maneira tradicional. Os autores mediram o solo em dois pântanos salgados escoceses. Abordagens de amostragem para predizer o SOCS através de algoritmos de AM também são apresentadas em Zeraatpisheh et al. [2021]; Ferreira et al. [2023]; Barragán-Pazmiño et al. [2024]. Sarkar and Somenahally [2023] exploraram as perspectivas na forma de modelos de AM no agroecossistema para entender melhor o processo de retirada de carbono do solo e Xie et al. [2024] forneceram informações detalhadas sobre locais de amostragem, profundidades, conteúdo de matéria orgânico do solo (MOS), densidade aparente (DB) e outras

propriedades físico-químicas do solo.

RN é um dos algoritmos que apresenta uma dificuldade de interpretabilidade e alguns artigos que usaram esse algoritmo juntamente com a RF foram encontrados. dos Santos et al. [2022] construíram modelos para quantificar o conteúdo de carbono total em solos no sul do Brasil, onde a RF foi considerado o melhor modelo por possuir melhor predição. Em Emadi et al. [2020]; Taghizadeh-Mehrjardi et al. [2020], o algoritmo de RN superou outros algoritmos de AM na previsão do SOCS. Temos também os trabalhos de Hateffard et al. [2023]; Zhang et al. [2023a]; Radocaj et al. [2024]; Fu et al. [2024]; Abbaszad et al. [2024], e o trabalho de Carbajal et al. [2024] que comparou as variáveis-alvo de carbono do solo entre os usos mais importantes da terra e analisou o desempenho de alguns métodos de ML para modelagem preditiva nas Terras Altas dos Andes Centrais do Peru.

Entre os artigos que conseguiram interpretar o modelo, já citados anteriormente, temos o Ross et al. [2020] que estabeleceu uma rede de monitoramento em todo o sudeste dos EUA para refinar a compreensão do armazenamento e dinâmica de carbono em florestas. Em Wadoux et al. [2020a] é relatado o desenvolvimento de um método estatístico chamado *Shapley* para interpretação de modelos complexos. Tal método foi implementado para o estudo da variação de estoque SOCS. O outro artigo que provê uma interpretação é o apresentado por Ferreira et al. [2023], como já explicado, serviu para predizer o SOCS em uma área remota da Floresta Amazônica.

Os incêndios nas matas são o principal agente de perturbação, muitas vezes limitando o acúmulo de camadas orgânicas, assim como a erosão no solo. Dois artigos abordaram esses temas. Hanes et al. [2022] desenvolveram mapas de espessura da camada orgânica e carga de combustível usando abordagens de AM para a região florestal da província de Alberta, Canadá. Compreender a distribuição espacial desses atributos é importante para melhorar as predições do comportamento do fogo. Aqdam et al. [2022], realizaram a predição, através de AM, das classes baixa e alta do fator de erodibilidade do solo (fator K) no noroeste do Irã.

Foram encontrados também artigos que abordaram assuntos diversos, como dados hipotéticos, série temporal (ST), modelagem aditiva generalizada (GAM), regressão Bayesiana, desenvolvimento de zonas de manejo do solo (MZs) e sensoriamento remoto em AGB (*Above Ground Biomass*). Wadoux et al. [2020b] estudaram como pseudo covariáveis podem ser usadas para predizer com acurácia, SOCS em um estudo de caso hipotético, usando AM. Os autores Schwab et al. [2022] investigaram, com base em uma ST de 40 meses, a dependência da hidrologia e da sazonalidade na descarga de sedimentos e carbono orgânico em uma pequena Bacia subalpina suíça. Mishra et al. [2022b] usaram duas abordagens de AM (RF e GAM, que é raramente encontrada na literatura) para selecionar preditores ambientais importantes de estoques SOCS. Já os autores Molla et al. [2022], avaliaram a capacidade do método geoestatístico avançado de predição empírica de EBKRP e RF para predizer e mapear o conteúdo de elementos potencialmente tóxicos (PTEs) do solo. Ma-

leki et al. [2023] investigaram MZs delineado com base nas diferentes propriedades do solo usando métodos de AM. Os autores Chan et al. [2023] combinaram sensoriamento remoto em AGB e o SOCS para melhorar a estimativa de carbono total.

Os trabalhos que tiveram como objetivo desenvolver um método de AM para predizer a MOS, como nitrogênio do solo, potássio, carbono orgânico, entre outros, foram Ye et al. [2021]; Heil et al. [2022]; Nia et al. [2022]; Adeniyi et al. [2023].

O que apresenta uma abordagem de ciência de dados que aglomera o espaço de parâmetros do solo em um número limitado de unidades funcionais de processamento do solo (SPUs) que podem ser usadas para executar modelos de processos agrícolas, temos o trabalho de Lieb [2022].

Teve um trabalho de Möller et al. [2022] que usou refletância de solo multitemporal (SRC) para predição SOCS através da modelagem da RF. Para quem usou os algoritmos de mínimos quadrados parciais (PLS), RF e SVM nas predições SOCS e Nitrogênio temos Rameurez et al. [2022].

Temos o trabalho de Takoutsinga and Heuvelink [2022] que comparou KR e RF para quantificar as incertezas das predições, ao mesmo tempo em que contabiliza erros de medição nos dados do solo. Os autores Hounkpatin et al. [2021] avaliaram como os modelos de AM globais e locais diferem para predizer a camada de húmus, solo mineral e estoque total de SOCS nos ecossistemas florestais da Suécia. E também Westhuizen et al. [2023], que através do banco de dados de solo LUCAS da Eurostat, aplicaram modelo RF para mapear SOCS e o Nitrogênio.

Por fim, temos outros trabalhos selecionados que utilizaram RF. Os trabalhos dos autores Zhang et al. [2023b,a]; Broeg et al. [2023b]; Yu et al. [2024]; Chen et al. [2024]; Nwaogu et al. [2024] que estimaram SOCS, os trabalhos de Chinilin and Savin [2023]; Agaba et al. [2024] que mapearam SOCS, e também, o trabalho de Wang et al. [2024b] que utilizou as duas técnicas, estimação e mapeamento do SOCS, através da análise de agrupamento multivariada e AM.

III.12.2 Artigos que não usaram RF

Para começar, temos os trabalhos que utilizaram regressão de SVM para predição de SOCS. O trabalho de Rengma et al. [2020] foi conduzido em uma bacia hidrográfica indiana do Himalaia, para mapear SOCS e frações texturais (areia, silte e argila). Temos também os autores Xu et al. [2021], que compararam métodos estatísticos multivariados como a Regressão de Mínimos Quadrados Parciais (PLSR), com os algoritmos de AM da SVM e RN Wavelet para estimativa de SOCS. Os autores Huang et al. [2023] decomporam os dados espectrais do solo usando transformada contínua de Wavelet e extrair as informações para detectar o conteúdo SOCS. O trabalho de Ravaioli et al. [2023] desenvolveu modelo baseado em pastagens permanentes semeadas, ricas em leguminosas (SBP) em Portugal e Ding et al. [2023] estimou a distribuição espacial do SOCS no Planalto Tibetano na China usando imagens de veículos aéreos não tripulados. Por fim, temos

dos autores Xiao et al. [2024] que compararam o modelo SVM com PLSR e RLM, dos Mango et al. [2024] que compararam com RN e PLRS e dos autores Zhang et al. [2024] que também compararam o SVM com o PLRS.

Kumara et al. [2012]; Liu et al. [2020] propuseram a Krigagem de Regressão Geograficamente Ponderada (GWRK) para um modelo de predição de SOCS. Já Bahri et al. [2022]; Zhang et al. [2023c] propuseram o algoritmo Floresta de Regressão Quantílica (QRF) para mapear estoques de SOCS.

Foram encontrados trabalhos que realizaram amostragem do solo para analisar propriedades do solo potencialmente indicativas de mudanças induzidas pelo fogo. Brook et al. [2022] examinaram o uso de imagens multiespectrais e de infravermelho, coletadas por sistemas aéreos não tripulados para determinar a gravidade do incêndio de acordo com a cor e as propriedades químicas das cinzas da vegetação. Os autores Beltrán-Marcosa et al. [2023] analisaram o diâmetro ponderado médio (MWD), teor de umidade do solo (SMC) e SOCS. Uma técnica de nitidez de imagem *Gram-Schmidt* (GS) foi utilizada.

Referente a outros trabalhos que utilizaram AM, temos os que modelaram através do algoritmo de RC, como o trabalho de Adhikari et al. [2020] que analisou o comportamento de dimensionamento do SOCS nos Estados Unidos. E também de Nawar et al. [2020], realizado para remover o efeito negativo de teor de umidade do solo na predição do SOCS, e dos autores John et al. [2021] que aplicaram três modelos cubistas diferentes para estimar o SOCS. Os autores Morais et al. [2021] utilizaram RN para estimativa rápida de SOCS. E por fim, o trabalho de Kanari et al. [2022] demonstrou um modelo de AM baseado em *Rock-Eval* que particiona de forma otimizada os SOCS ativos e estáveis.

Trabalhos que utilizaram funções de Pedotransferência (PTF) também foram encontrados. O trabalho de Ceddia et al. [2016] teve como objetivo predizer a densidade do solo (DS) a partir de propriedades do solo na região amazônica brasileira. E também o trabalho de Szatmari et al. [2023] que apresentou uma PTF de base cubista para predizer e mapear o conteúdo SOCS saturado dos solos superficiais na Hungria.

Ainda em relação ao armazenamento de carbono orgânico, foi recuperado o trabalho de Yang et al. [2022b] que explorou o efeito de longo prazo das mudanças no uso da terra em carbono do solo e apresentou um plano de uso sustentável da terra para abordar uma série de questões ambientais. Uma floresta de regressão quantílica (QRF) e modelo de aumento de gradiente estocástico (GBM) foram utilizados para predição do SOCS em Paulo et al. [2022].

Temos o trabalho de Kim et al. [2022] que investigou as características espectrais do solo superficial de quatro grandes rios na República da Coreia e desenvolveu um modelo de predição de conteúdo da MOS baseado em AM. Temos o trabalho de Kebonye et al. [2023] que comparou as

predições SOCS para a República Tcheca enquanto variavam os modelos CART com base em estratégias de divisão de dados. E também em Emamgholizadeh et al. [2023], os autores estudaram as propriedades físicas e químicas do solo em um sistema de inferência Fuzzy baseado em rede adaptativa (ANFIS), do modelo ANFIS acoplado por evolução diferencial (ANFIS-DE) e do modelo ANFIS acoplado por otimização de partículas (ANFIS-PSO), que foram usados para a predição da capacidade de troca catiônica (CTC).

Por fim, temos os trabalhos de Lopatka et al. [2023] que fez a predição de SOCS utilizando método de AM de regressão linear regularizada (LASSO), de Gouda et al. [2024] que mapeou SOCS utilizando RN e Regressão de Processo Gaussiano (GPR) e Lei et al. [2024] que estudou a distribuição espacial e desenvolveu o modelo XGBoost para a predição de SOCS.

III.13 Artigo referência para comparação dos resultados

O trabalho de Ceddia et al. [2016] foi o que mais se aproximou do tema deste trabalho entre os 126 artigos selecionados. A Tabela III.8 mostra as semelhanças e diferenças entre o trabalho proposto nesta dissertação e o trabalho de Ceddia et al. [2016]. Em ambos os trabalhos, usam atributos do solo, estimam SOCS na região da Amazônia e calculam o indicador de acurácia MAE. As diferenças encontradas são que o trabalho de Ceddia et al. [2015] utiliza o modelo de Regressão Linear Múltipla (RLM) e não o AM, para a seleção de atributos utiliza o Critério de Informação Akaike (AIC) e não a Correlação de Pearson e o método *Wrapper*.

O trabalho de Ceddia et al. [2016] não utilizou a técnica de otimização de hiperprâmetros no modelo. Foram desenvolvidas funções de Pedotransferência (PTF) que são usados como modelos físico-matemáticos que permitem a estimativa de SOCS a partir de dados do solo.

Tabela III.8: Comparação dos resultados.

	Ceddia et al. [2016]	**Este trabalho**
Aprendizado de Máquina		X
Modelo de Regressão Linear Múltipla	X	
Funções de Pedotransferência	X	
Atributos do solo	X	X
Correlação de Pearson		X
Método Wrapper		X
Critério de Informação Akaike	X	
Região da Amazônia	X	X
SOCS = DS*MOS*Espessura	X	X
Desempenho MAE	X	X
Otimização Hiperparâmetros		X

Capítulo IV Metodologia

Este capítulo apresenta o fluxo de trabalho (*Pipeline*) da metodologia desta pesquisa, com uma visão geral das etapas desenvolvidas neste trabalho, os hiperparâmetros determinados para os algoritmos, uma breve descrição sobre o problema de determinação de Estoque de Carbono (SOCS), bem como da base de dados utilizada para treinar e validar os modelos de AM. Por fim, a análise descritiva dos dados através de gráficos box plot, onde possível visualizar a distribuição de dados com base em cinco medidas estatísticas.

IV.1 Fluxo de Trabalho

O *pipeline* da metodologia iterativa implementada neste trabalho é apresentado na Figura IV.1, que fornece uma visão geral das etapas fundamentais do processo. Essas etapas incluem: *(i)* Base de Dados, *(ii)* Pré-processamento, *(iii)* Seleção de Atributos, *(iv)* Treinamento do Modelo, *(v)* Ajuste de Hiperparâmetros e *(vi)* Avaliação de Desempenho.

O ponto de partida no ciclo de desenvolvimento de qualquer projeto de Aprendizado de Máquina (AM) é a preparação dos dados, que envolve a criação da variável-alvo (SOCS) e a seleção de atributos relevantes. Para isso, foram empregadas técnicas como a Correlação de Pearson e métodos baseados na abordagem *Wrapper*. Com os dados devidamente preparados, a próxima etapa consiste em aplicar técnicas de otimização de hiperparâmetros para garantir que os modelos de AM sejam ajustados de maneira eficiente.

Nos algoritmos de seleção de atributos (*Wrapper*) e de otimização (*Random Search*, *Successive Halving* e *Grid Search*), foram definidos no hiperparâmetro da validação cruzada *cv*, o *K-fold* igual a 5.

Após esse ajuste, os algoritmos de AM são treinados e avaliados com base nas métricas de desempenho, permitindo identificar e selecionar o melhor modelo para o problema aqui abordado. A execução correta e rigorosa de cada uma dessas etapas desde o início é essencial para garantir resultados precisos e confiáveis, destacando a importância da abordagem estruturada e iterativa desenvolvida.

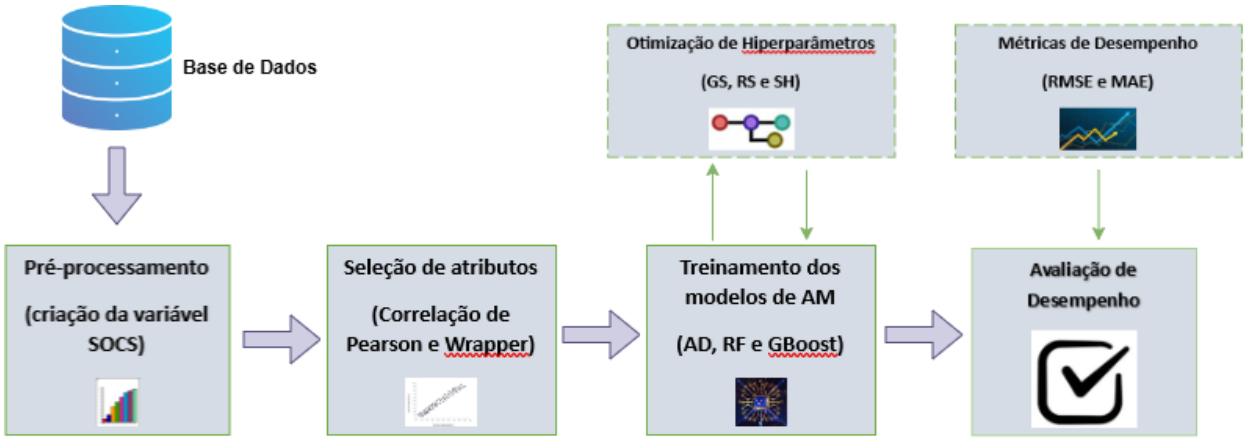


Figura IV.1: *Pipeline* da metodologia implementada. Adaptado de: [Temizel et al., 2020].

IV.2 Os hiperparâmetros definidos

A seguir são apresentados os hiperparâmetros definidos nos algoritmos dos métodos *Wrapper* na seleção de atributos e nos algoritmos na aplicação da otimização nos modelos de AM.

IV.2.1 Métodos *Wrapper*

Antes da aplicação dos métodos *Wrapper*, foram definidos os parâmetros do EFS[3] e do SFS[4], como mostra a Tabela IV.1.

Tabela IV.1: Métodos Wrapper.

Variável alvo	Método Wrapper	Parâmetros
SOCS	EFS	n_estimators = 100 random_state = 0 min_features = 1 max_features = 7 scoring = MAE ou RMSE cv = 5
SOCS	SFS	n_estimators = 100 random_state = 0 n_features_to_select = “auto” tol=0.1, 0.01 ou 0.001 direction=’forward’ ou ’backward’ scoring = RMSE ou MAE cv = 5

Onde:

n_estimators: é o número de árvores na floresta aleatória.

random_state: é o estado do gerador de números pseudoaleatórios usado para amostragem

[3] `https://rasbt.github.io/mlxtend/user_guide/feature_selection/ExhaustiveFeatureSelector/`

[4] `https://scikit-learn.org/dev/modules/generated/sklearn.feature_selection.SequentialFeatureSelector.html`

aleatória uniforme de listas de valores possíveis. Esse parâmetro serve para que seu código possa ser reproduzido novamente gerando os mesmos resultados, garantindo a reprodutibilidade do processo.

scoring: é a estratégia para avaliar o desempenho do modelo de validação cruzada no conjunto de testes.

cv: determina a estratégia de divisão de validação cruzada.

min_features: é o número mínimo de atributos que um subconjunto pode ter.

max_features: é o número máximo de atributos que um subconjunto pode ter.

n_features_to_select: número de atributos em um subconjunto aleatoriamente.

tol: se a pontuação não aumentar pelo menos o valor do tol, entre duas adições ou remoções consecutivas de atributos, o processo de adição ou remoção de atributos é interrompido. O tol está habilitado somente quando n_features_to_select = "auto". As métricas de avaliação utilizadas são MAE ou RMSE. Um parâmetro é usado para definir o limite no qual a pesquisa será interrompida. Neste projeto, foram testados três valores para tol (0.1, 0.01 e 0.001), o que significa que, se a métrica de desempenho aumentar acima de tol, a pesquisa será encerrada.

direction: observando que temos uma malha de valores possíveis, este parâmetro realiza uma busca destes valores para frente (*forward*) ou para trás (*backward*) nesta malha de possibilidades.

IV.2.2 Otimização por hiperparâmetros

Antes da aplicação dos hiperparâmetros na otimização, foram definidos os hiperparâmetros do RS[5], SH[6] e do GS[7], como mostra a Tabela IV.2.

Tabela IV.2: Hiperparâmetros definidos.

Variável alvo	**Otimização de Hiperparâmetro**	**Hiperparâmetros**
SOCS	RS	param_distributions = dict( ) n_iter = 10 a 500 scoring = RMSE ou MAE n_jobs = 1 cv = 5 random_state = 0
SOCS	SH	param_distributions = dict( ) scoring = RMSE ou MAE n_jobs = 1 cv = 5 random_state = 0
SOCS	GS	param_distributions = dict( ) scoring = RMSE ou MAE n_jobs = 1 cv = 5

[5] `https://scikit-learn.org/1.5/modules/generated/sklearn.model_selection.RandomizedSearchCV.html`
[6] `https://scikit-learn.org/dev/modules/generated/sklearn.model_selection.HalvingGridSearchCV.html`
[7] `https://scikit-learn.org/dev/modules/generated/sklearn.model_selection.GridSearchCV.html`

Onde:

param_distributions: está relacionado à função *dict( )*, usada para criar e manipular dicionários com nomes e valores dos parâmetros, como por exemplo, os desempenhos utilizados neste trabalho. Cria, uniformemente, uma lista de parâmetros. Uma distribuição uniforme assume que todas as amostras dos parâmetros sejam igualmente prováveis.

n_iter: número de configurações de parâmetros amostradas. Negocia tempo de execução versus qualidade da solução. No código foram testados cinco valores para n_iter (10, 50, 150, 250 e 500).

scoring: é a estratégia para avaliar o desempenho do modelo de validação cruzada no conjunto de testes.

n_jobs: número de trabalhos a serem executados em paralelo. Nenhum significa valor 1, a menos que esteja em um contexto.

cv: determina a estratégia de divisão de validação cruzada.

random_state: estado do gerador de números pseudoaleatórios usado para amostragem aleatória uniforme de listas de valores possíveis. Esse parâmetro serve para que seu código possa ser reproduzido novamente gerando os mesmos resultados, seja em outro momento, ou por outra pessoa, ou em outro computador.

IV.3 Determinação do Estoque de Carbono Orgânico do solo

Uma forma de determinar o Estoque de Carbono Orgânico do Solo (SOCS) consiste em somar os teores de matéria orgânica (MOS) por horizonte, sendo calculado por meio do produto da densidade aparente do solo (DS) pelo MOS e pela profundidade do horizonte (Espessura) [Ferreira et al., 2023], conforme descrito na Equação IV.1. Esta medida fornece a massa de carbono por unidade de área.

$$SOCS = DS * MOS * Espessura \tag{IV.1}$$

Onde, SOCS em Kg C m^{-2}, MOS em g Kg^{-1}, DS em Mg m^{-3} e Espessura em m.

IV.4 Base de Dados

O conjunto de dados utilizado nesta pesquisa é baseado nos trabalhos anteriores realizados por Haddad et al. [2017, 2018], tendo sido aprimorado com informações adicionais sobre as regiões geográficas, seguindo a definição do Instituto Brasileiro de Geografia e Estatística (IBGE). O conjunto de dados completo consiste de 3404 registros, abrangendo todas as cinco regiões geográficas do Brasil (Norte, Nordeste, Centro-Oeste, Sudeste e Sul).

A Figura IV.2 mostra a distribuição dos dados no território brasileiro separados pelos biomas

brasileiros, que são constituídas por um complexo conjunto de ecossistemas, apresentando características semelhantes como vegetação, fauna, solo e clima.

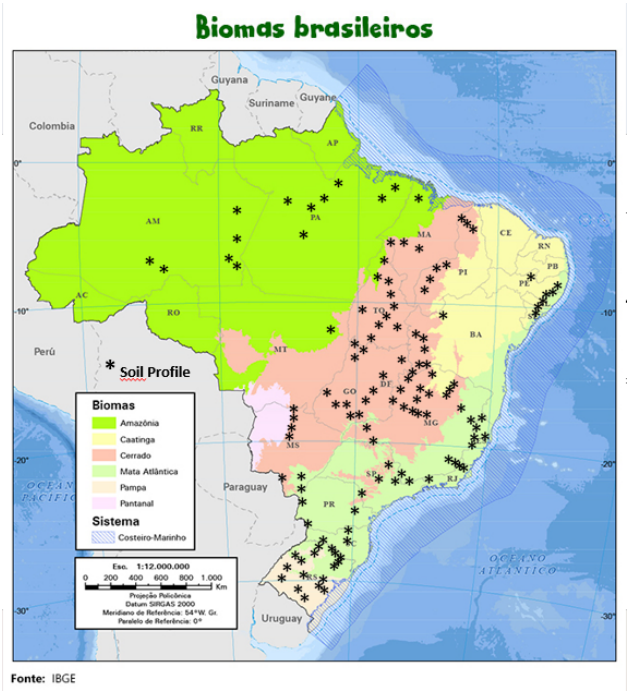


Figura IV.2: Mapa com a localização dos pontos amostrados. Adaptado de: [Haddad et al., 2018].

Para fins de comparação dos resultados com o trabalho Ferreira et al. [2023], foi considerado somente os dados da região da Amazônia. Através de uma filtragem da variável relacionada à localização na base de dados completa, foram extraídas 479 amostras referente à Amazônia obtidas junto ao Departamento de Solos da Universidade Federal Rural do Rio de Janeiro (UFRRJ). A Figura IV.3 mostra a área da Amazônia no Brasil.

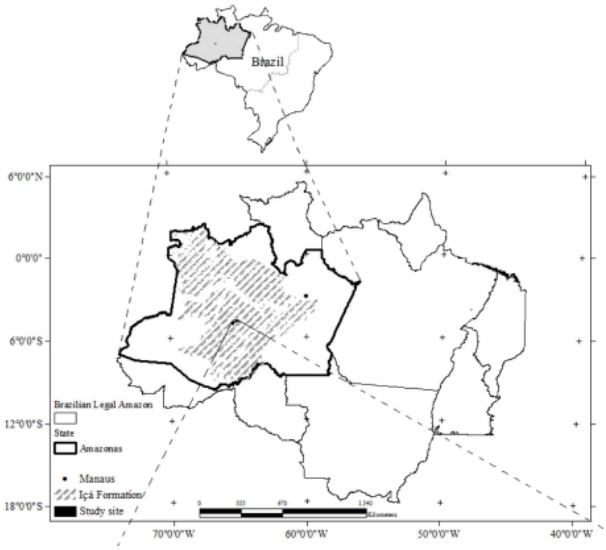


Figura IV.3: Localização da Amazônia - Brasil. Fonte: [Ceddia et al., 2015].

As variáveis apresentadas na base de dados que foram utilizadas neste trabalho, com a inclusão da variável SOCS, são descritas na Tabela IV.3. A variável *Espessura* foi calculada através da diferença das profundidades iniciais e finais em relação à superfície com marco inicial (zero metro). As profundidades foram extraídas na base de dados completa, onde a profundidade inicial se refere à variável *hzdepup*, e a profundidade final, à variável *hzdeplow*. A variável *MOS* está relacionada ao teor de carbono orgânico no solo.

Tabela IV.3: Variáveis utilizadas. Adaptado de: [Haddad et al., 2018, 2017].

Variável	**Especificação**
$SOCS$	Estoque de carbono no solo
DS	Densidade do solo
$Sand$	Proporção de partículas de areia no solo
$Silte$	Proporção de partículas de silte no solo
$Clay$	Proporção de partículas de argila no solo
pH	pH do solo em água
SB	Teor de saturação por bases no solo
Al	Teor de alumínio no solo
H	Teor de hidrogênio no solo
MOS	Teor de matéria orgânica no solo
$Espessura$	Espessura do solo

A Tabela IV.4 apresenta um recorte com as primeiras linhas do conjunto de dados utilizado considerando a região da Amazônia [Haddad et al., 2018] de um total de 479 linhas.

Tabela IV.4: As primeiras observações da base de dados.

	SOCS	**DS**	**Sand**	**Silte**	**Clay**	**pH**	**SB**	**Al**	**H**	**MOS**	**Espessura**
1	1,1868	0,92	439	434	127	36	25,2	47,0	124,0	21,5	0,06
2	2,3509	1,37	370	470	164	43	18,2	25,0	60,0	7,8	0,22
3	2,1978	1,49	318	331	345	48	16,1	43,0	17,0	5,9	0,25
4	2,2623	1,47	280	320	401	53	10,0	50,0	21,0	5,7	0,27
5	1,6799	0,61	260	422	319	37	30,2	52,5	168,5	30,6	0,09
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	
479	1,0592	0,91	385	556	59	42	27,0	20,0	165,0	19,4	0,06

IV.5 Análise descritiva dos dados

A análise descritiva dos dados foi feita através do gráfico box plot que é uma ferramenta gráfica utilizada para representar um conjunto de dados distribuídos em cinco diferentes medidas (mínimo, primeiro quartil (Q1), mediana (Q2), terceiro quartil (Q3) e máximo), que possibilitam diversos tipos de observações em uma única análise.

A Figura IV.4 mostra o gráfico box plot da variável alvo SOCS calculada pela fórmula IV.1. E a Figura IV.5 mostra o gráfico box plot dos atributos da base de dados.

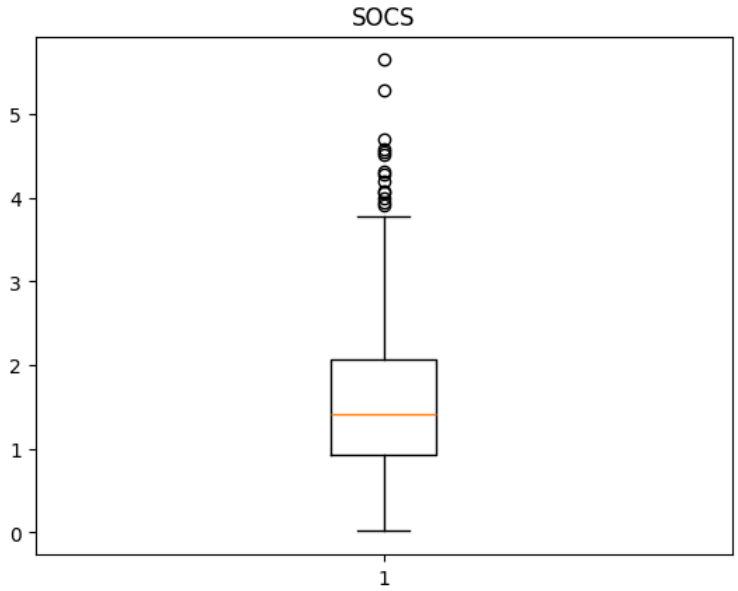


Figura IV.4: Box plot - SOCS.

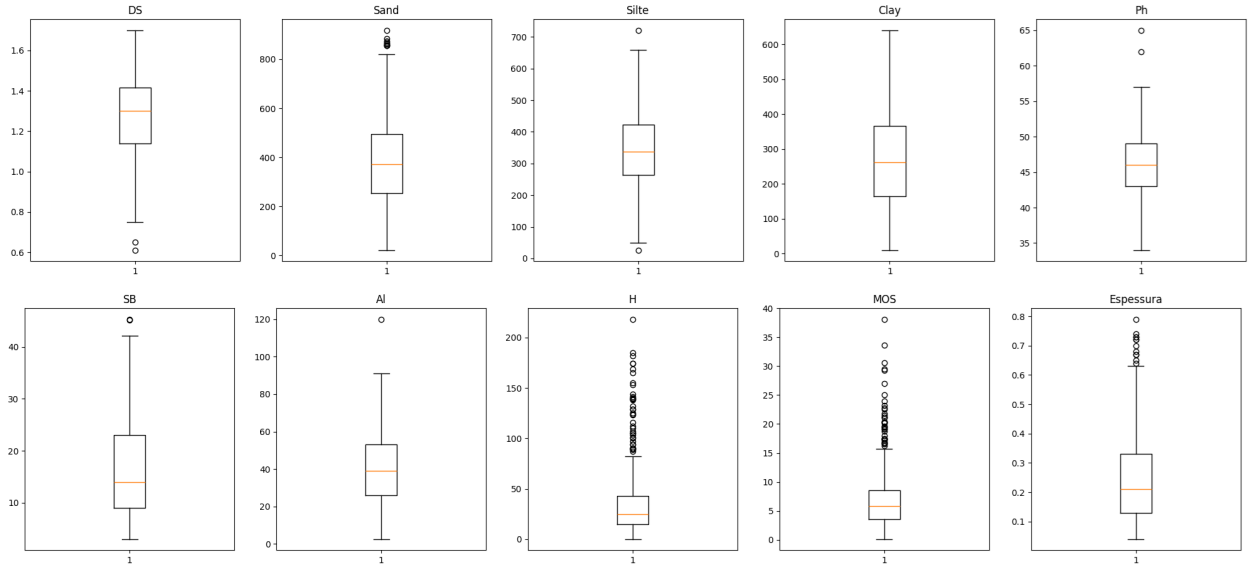


Figura IV.5: Box plot - Atributos.

A Tabela IV.5 apresenta a análise descritiva dos dados com base em um resumo de cinco números do box plot, para a variável alvo SOCS e os 10 atributos da base de dados.

Tabela IV.5: Análise descritiva dos dados.

	count	**mínimo**	**Q1**	**mediana**	**Q3**	**máximo**
SOCS	479	0,0164	0,9345	1,4100	2,0752	5,6445
DS	479	0,61	1,14	1,30	1,42	1,70
Sand	479	22	255	373	495	918
Silte	479	26	264	338	422	721
Clay	479	10	164	262	366	640
pH	479	34	43	46	49	65
SB	479	3	9	14	23	45
Al	479	2,5	26	39	53	120
H	479	0	15	25	43	218
MOS	479	0,10	3,57	5,80	8,50	38,10
Espessura	479	0,04	0,13	0,21	0,33	0,79

Pelos gráficos box plot das Figuras IV.4 e IV.5, os atributos SOCS, Sand, H, MOS e Espessura apresentam *outliers*, que são valores atípicos ou discrepantes que se afastam dos outros valores de um conjunto de dados. Esses *outliers* não foram excluídos para que não ocorram perdas de informações e prejudiquem as análises dos modelos. Nesse caso, a mediana é uma boa medida pois é menos afetada pelos *outliers*, por ser o valor central de um conjunto de dados ordenado. Ao contrário da média, que é sensível aos *outliers*, pois um único valor extremo pode aumentar ou diminuir significativamente.

A Figura IV.4 e a Tabela IV.5 mostram os resultados da análise descritiva da variável SOCS. Analisando o box plot e a tabela, temos as seguintes conclusões referente à variável alvo:

- A amplitude da variável SOCS, que é a diferença entre o maior e o menor valor do conjunto, é igual a 5,6281.
- O 1º quartil ou 25º percentil do SOCS é de valor 0,9345. Isso significa que pelo menos 25% dos valores de SOCS tem no máximo valor igual a 0,9345.
- O 3º quartil ou 75º percentil do SOCS é de valor 2,0752. Isso significa que pelo menos 75% dos valores de SOCS tem no máximo valor igual a 2,0752.
- Como a mediana está mais próxima do 1º quartil, os dados têm uma distribuição assimétrica positiva. Isso significa que existe uma maior concentração de valores do SOCS na zona de valores inferiores da base de dados.

Analogamente, essas análises podem ser realizadas para todos os outros dez atributos, permitindo que conclusões semelhantes sejam obtidas.

Capítulo V Resultados

Este capítulo apresenta os resultados da metodologia proposta para resolver o problema de predição do estoque de carbono orgânico no solo (SOCS). Os resultados foram comparados com o trabalho de Ceddia et al. [2016], o que mais se aproximou do tema deste trabalho, como citado na Seção III.13. Foram comparadas as técnicas de RF, AD e GBoost, usando como métricas de desempenho, o RMSE e o MAE, com e sem a estratégia de otimização de hiperparâmetros, para verificar se a técnica de otimização pode auxiliar no desenvolvimento de modelos mais precisos.

Os experimentos foram realizados com rotinas computacionais implementadas em Python versão 3 na plataforma *Google Colaboratory*, em uma máquina Dell Inspiron Intel Core i7 − 1255U CPU 3.20GHz, com 10 núcleos, 16GB de memória RAM e sistema operacional Windows 11. As bibliotecas utilizadas foram *pandas*[8] versão 1.5.3, *numpy*[9] versão 1.24.2 e *scikit-learn*[10] versão 1.2.2.

V.1 Resultados da seleção de atributos

Esta seção apresenta os resultados referentes à Correlação de Pearson e aos métodos *Wrapper*. Pela Equação IV.1, o valor do SOCS foi calculado em função dos atributos DS, MOS e Espessura. Para evitar uma redudância, esses três atributos não foram incluídos na seleção.

V.1.1 Correlação de Pearson

A Tabela V.1 e a Figura V.1 apresentam as correlações das variáveis utilizadas. As cores que indicam o grau de correlação no mapa de calor, apresentadas na Figura V.1, podem ajudar na visualização dos dados. Sendo que valores mais vermelhos indicam uma forte correlação positiva, enquanto valores mais azuis indicam uma forte correlação negativa. Enquanto os valores próximos de 0 indicam uma correlação fraca, representados pelas cores laranja claro, azul claro e amarelo.

Observando os resultados da tabela, verificam-se que as correlações dos atributos em relação à variável de interesse SOCS têm valores fracos, todos menos de $0,30$. Isso pode ser verificado no mapa de calor, onde aparecem somente as cores de intensidade baixa (amarelo e azul claro).

Já os resultados das correlações entre os atributos, no mapa de calor, a maioria tem cores

[8] https://pandas.pydata.org/
[9] https://numpy.org/
[10] https://scikit-learn.org/stable/

de intensidade baixa, que representam os valores mais fracos. Entre os valores fortes, aparece o azul escuro que é a maior correlação negativa $(-0,71618)$, referente aos atributos Clay e Sand. E também, um laranja avermelhado que é a maior correlação positiva $(0,67843)$, referente aos atributos Al e Clay.

Tabela V.1: Matriz de correlações das variáveis.

	SOCS	**Sand**	**Silte**	**Clay**	**pH**	**SB**	**Al**	**H**
SOCS	1,000000	-0,205174	-0,015289	0,281046	0,021093	-0,120637	0,216889	-0,184446
Sand	-0,205174	1,000000	-0,642084	-0,716189	-0,103788	0,316011	-0,582277	0,171922
Silte	-0,015289	-0,642084	1,000000	-0,075162	-0,087227	-0,011741	0,086544	0,142578
Clay	0,281046	-0,716189	-0,075162	1,000000	0,214368	-0,400301	0,678433	-0,353291
pH	0,021093	-0,103788	-0,087227	0,214368	1,000000	-0,339582	-0,037589	-0,436632
SB	-0,120637	0,316011	-0,011741	-0,400301	-0,339582	1,000000	-0,262671	0,437423
Al	0,216889	-0,582277	0,086544	0,678433	-0,037589	-0,262671	1,000000	-0,188689
H	-0,184446	0,171922	0,142578	-0,353291	-0,436632	0,437423	-0,188689	1,000000

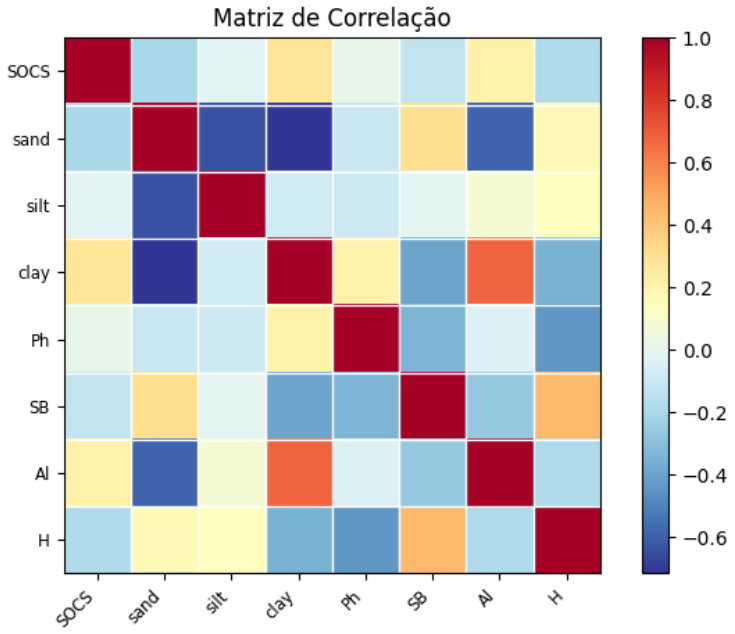


Figura V.1: Mapa de calor com a Correlações das variáveis.

A Tabela V.2 apresenta em ordem decrescente, os resultados em módulo das correlação da variável alvo SOCS com os sete atributos, e também, os resultados em módulo das maiores correlações entre os atributos (variáveis independentes).

Tabela V.2: Comparação das correlações das variáveis.

SOCS × atributos	**atributos × atributos**
SOCS × Clay (0,281046)	Clay × Sand (0,71618)
SOCS × Al (0,216889)	Al × Clay (0,67843)
SOCS × Sand (0,205174)	Silte × Sand (0,642084)
SOCS × H (0,184446)	Al × Sand (0,582277)
SOCS × SB (0,120637)	H × SB (0,437423)
SOCS × pH (0,021093)	H × pH (0,436632)
SOCS × Silte (0,015289)	H × Clay (0,353291)

Avaliando a correlação linear dos atributos com a variável SOCS percebe-se que os atributos Clay, Al e Sand apresentam maior correlação linear com a variável de interesse (maiores que $0,20$). São os atributos mais dependentes da variável SOCS, influenciando em uma possível predição, sendo, desta forma, mais relevantes para a criação de um eventual modelo preditivo. A variável Silte foi que apresentou a menor correlação linear, conduzindo a compreensão de que ela não é relevante para o modelo preditivo do SOCS.

A baixa correlação do Silte era um resultado esperado, pois conforme a matriz de correlação do artigo Haddad et al. [2017] nota-se resultados semelhantes a esse comportamento. No trabalho de Haddad et al. [2018], através da matriz de correlação, também foi encontrado uma baixa correlação do Silte em relação à densidade.

Avaliando a correlação linear entre os atributos, as variáveis Clay e Sand apresentam a maior correlação entre as variáveis independentes. Tal resultado indica que o modelo a ser desenvolvido pode utilizar somente uma das duas variáveis.

Esse resultado também era esperado. Nos trabalhos Haddad et al. [2017, 2018] os resultados encontrados apresentaram alta correlação entre as variáveis Clay e Sand.

V.1.2 Métodos *Wrapper*

No método *Wrapper* SFS, testamos para o hiperparâmetro *tol* os valores {0,1; 0,01; 0,001} nas duas direções e não houve diferença nos resultados dos subconjuntos ideais.

As Tabelas V.3 e V.4 mostram os resultados de EFS e SFS nas direções *forward* e *backward*.

Tabela V.3: Resultados *Wrapper - EFS*.

Métrica de Desempenho	***Wrapper* - EFS**
	Subconjunto ideal
RMSE	Sand, Silte, SB, Al e H
MAE	Sand, Silte, PH, SB, Al e H
	Atributos excluídos
RMSE	Clay e PH
MAE	Clay

Tabela V.4: Resultados *Wrapper - SFS*.

Métrica de Desempenho	***Wrapper* - SFS direction='forward' tol=0.1, 0.01 ou 0.001**	***Wrapper* - SFS direction='backward' tol=0.1, 0.01 ou 0.001**
	Subconjunto ideal	**Subconjunto ideal**
RMSE	pH	Sand, Silte, pH, SB, Al e H
MAE	pH	Sand, Silte, pH, SB, Al e H
	Atributos excluídos	**Atributos excluídos**
RMSE	Sand, Silte, Clay, SB, Al e H	Clay
MAE	Sand, Silte, Clay, SB, Al e H	Clay

A Tabela V.3 mostra que, no método de seleção de atributos *Wrapper* - EFS, a maioria dos atributos está no subconjunto ideal e é relevante para o modelo preditivo, nas duas métricas de desempenho RMSE e MAE.

A Tabela V.4, nas duas métricas de desempenho, os resultados mostram que o método *Wrapper* - SFS *forward* possui somente uma variável relevante para o modelo (pH), e o método *Wrapper* - SFS *backward*, somente uma variável não foi relevante para o modelo preditivo (Clay).

V.1.3 Observações referentes aos resultados da seleção de atributos

Pelos resultados da Tabela V.2, na Correlação de Pearson, o Silte foi o atributo menos relevante para o modelo preditivo. O Clay aparece como a maior correlação com a variável SOCS, porém o valor 0,281046 está muito distante do ideal. O valor ideal da correlação foi explicado na Seção II.5, que seria um valor próximo a $1,0$. Assim, a correlação entre Clay e Sand apresenta o maior valor e uma delas pode ser excluída.

Pelos resultados das Tabelas V.3 e V.4, o atributo Clay não aparece no subconjunto ideal, em ambos os métodos *Wrapper*. Analisando os resultados da Correlação de Pearson e os métodos *Wrapper*, optamos que os atributos Silte e Clay poderiam ser excluídos e não são tão relevantes para o modelo preditivo.

V.2 Determinação do número de iterações ideal

Como consta na Tabela IV.2 da Seção II.6, o número de iterações (n_iter) foi testado em cinco valores distintos {10, 50, 150, 250 e 500} na técnica de otimização RS.

Em contraste com a técnica GS, na RS nem todos os valores de parâmetros são avaliados, mas sim um número fixo de configurações de parâmetros é amostrado. O número de iterações é crucial em algoritmos de otimização, onde a atualização dos parâmetros ocorre repetidamente até que um critério de parada seja atingido. Um número insuficiente de iterações pode resultar em um modelo subajustado, enquanto que um número excessivo pode levar ao sobreajuste [Alves, 2024].

O número de iterações é um conceito central em estatística, análise de dados e ciência de dados, influenciando diretamente a qualidade e a eficiência dos modelos. A escolha do número de iterações pode impactar diretamente a acurácia dos resultados obtidos e a eficiência do processo de análise [Alves, 2024].

Os critérios de parada são condições que determinam quando o processo de iteração deve ser interrompido, como por exemplo, onde as mudanças entre iterações são mínimas. A Tabela V.5 mostra que houve pouca diferença entre os resultados dos parâmetros nas cinco iterações testadas.

Como os resultados dos parâmetros nas cinco iterações foram próximos, a escolha pelo número de iterações ideal foi o maior valor, 500, para que tenhamos um desempenho geral mais sólido dos modelos de AM.

Tabela V.5: Número de iterações (RS).

Modelo de AM (Random Search)	**RMSE**	**MAE**	**n_iter**
RF (resultados)	0,939	0,714	10
	0,942	0,711	50
	0,941	0,711	150
	0,953	0,716	250
	0,938	0,715	500
AD (resultados)	1,381	1,049	10
	1,371	1,006	50
	1,375	1,032	150
	1,395	1,004	250
	1,368	1,027	500
GBoost (resultados)	0,965	0,725	10
	0,966	0,726	50
	0,962	0,726	150
	0,964	0,726	250
	0,964	0,725	500

V.3 Comparação com os resultados do artigo de Ceddia et al. [2016]

O modelo de RF gerado, que obteve o melhor desempenho entre os modelos de AM deste trabalho, foi comparado com o resultado do artigo Ceddia et al. [2016]. A comparação foi realizada utilizando a métrica MAE, empregada em ambos os trabalhos. A escolha deste artigo também levou em consideração que ambos os trabalhos utilizam atributos de solos para estimar SOCS, além de focarem na região da Amazônia Central, uma região de elevada importância ambiental e econômica para toda a sociedade.

A Tabela V.6 apresenta a comparação dos resultados do MAE obtidos pela técnica de RF (n_iter = 500) desenvolvidos aqui com o resultado do MAE obtido em Ceddia et al. [2016] para o SOCS. Observa-se que o resultado de Ceddia et al. [2016] alcançou um melhor valor para MAE. No entanto, apesar de terem obtidos valores melhores, o modelo linear descrito pelos autores apresentou um R^2

de aproximadamente $0,69$, o que demonstra que o modelo não consegue explicar cerca de 31% dos dados, indicando que novas abordagens precisam ser realizadas.

Neste contexto, o presente trabalho buscou realizar melhorias nos modelos inicialmente propostos. Para tanto, foi utilizada a estratégia de otimização dos hiperparâmetros, conforme será descrito na próxima Seção.

Tabela V.6: Comparação dos resultados.

Modelo	**Variável alvo**	**MAE**
RLM (Ceddia et al. [2016])	SOCS	0,090
RF	SOCS RS	0,715
	SOCS SH	0,706
	SOCS GS	0,718

V.4 Resultados da estimação de SOCS com Otimização de Hiperparâmetros

A Tabela V.7 apresenta os resultados, considerando as métricas de desempenho utilizadas para avaliação (RMSE e MAE), tanto com a utilização da técnica de otimização de hiperparâmetros quanto sem a técnica de otimização.

Analisando a métrica de MAE, o melhor valor $(0,706)$ foi obtido pelo modelo *Random Forest* (RF) com a técnica de otimização *Successive Halving* (SH). Esse valor indica que o modelo apresenta um bom desempenho na redução da média dos erros absolutos ao realizar predições.

Em relação ao RMSE, observamos o melhor valor $(0,938)$ também para o modelo RF, porém, agora com a técnica de otimização *Random Search* (RS).

Pelos resultados apresentados usando a estratégia de otimização de hiperparâmetros, o modelo de Aprendizado de Máquina que teve o melhor desempenho foi o *Random Forest*.

Comparando os resultados das três técnicas de otimização de hiperparâmetros com os resultados sem utilização de otimização, observa-se que a estratégia de otimização obtém melhores resultados para a métrica RMSE em todos os três modelos de AM.

Esses resultados indicam que um tratamento adequado dos dados com técnicas de pré-processamento e a estratégia de otimização dos hiperparâmetros podem auxiliar no desenvolvimento de modelos mais precisos. Cabe salientar que apesar desta melhoria, os resultados aqui obtidos ainda são inferiores ao de Ceddia et al. [2016], o que indica que novas abordagens devem ser avaliadas.

Tabela V.7: Estimação SOCS.

Variável alvo	**Modelo de AM**	**Hiperparâmetro**	**RMSE**	**MAE**
SOCS	RF	RS	0,938	0,715
		SH	0,944	0,706
		GS	0,947	0,718
		s/ otimização	0,985	0,734
SOCS	AD	RS	1,368	1,027
		SH	1,385	1,027
		GS	1,369	1,040
		s/ otimização	1,420	1,017
SOCS	GBoost	RS	0,964	0,725
		SH	0,964	0,726
		GS	0,964	0,727
		s/ otimização	0,976	0,730

Capítulo VI Considerações Finais

A determinação precisa do estoque de SOCS é essencial para a compreensão e mitigação das mudanças climáticas. O carbono armazenado no solo desempenha um papel fundamental no ciclo global do carbono, influenciando diretamente a concentração de dióxido de carbono na atmosfera. No Brasil, um país onde o setor agropecuário representa boa parte do PIB (7,4% do total no 1.º trimestre 2024)[11], compreender o comportamento do carbono no solo pode auxiliar em acordos internacionais, como o cumprimento dos Objetivos de Desenvolvimento Sustentável da ONU[12], especialmente o ODS 13 (ação contra a mudança global do clima) e auxiliar no crescimento econômico.

Neste contexto, o desenvolvimento de modelos de AM na área de solos pode auxiliar o país significativamente, pois tais modelos têm a capacidade de lidar com grandes volumes de dados e capturar relações complexas entre variáveis, permitindo o desenvolvimento de modelos mais precisos, ajudando em uma melhor compreensão dos solos brasileiros. No entanto, o desenvolvimento de modelos precisos requer uma otimização cuidadosa dos seus hiperparâmetros. Técnicas de otimização como *grid search*, *random search* e *successive halving* são essenciais para aprimorar esses modelos e alcançar predições mais precisas e confiáveis.

O presente trabalho demonstra a eficácia das técnicas de AM, especialmente RF, na estimativa do SOCS em solos brasileiros. Os resultados obtidos indicam que a seleção dos dados e otimização de hiperparâmetros feitos de forma adequada são cruciais para melhorar a acurácia dos modelos preditivos. Em particular, os modelos desenvolvidos superaram quase todas as abordagens sem utilizar técnicas de otimização.

Embora a técnica RF tenha mostrado bons resultados de predição, os modelos desenvolvidos são de difícil interpretação, mesmo com o pré-processamento realizado neste trabalho, onde foram identificadas as variáveis que mais contribuíram para estimar o valor de SOCS. Embora RF ofereça uma alta qualidade de predição, predições distorcidas ainda podem ocorrer e serem extremamente prejudiciais, comprometendo os resultados. Modelos de AM interpretáveis são essenciais para entender possíveis distorções e tendências e o seu uso aumentou significativamente [Vidal et al., 2020]. Neste contexto, para trabalhos futuros, propomos aprimorar as técnicas denominadas de árvores renascidas e procurar um regressor mais simples que reproduza fielmente o comportamento

[11] `https://cnabrasil.org.br/publicacoes/pib-da-agropecuaria-cresce-11-3-no-primeiro-tri-de-2024`
[12] `https://brasil.un.org/pt-br/sdgs`

de um conjunto de árvores preservando todas as características do modelo gerado, como realizado no trabalho apresentado em [Vidal et al., 2020].

Para trabalhos futuros, propomos testar outros algoritmos de AM e buscar regressores que possam facilitar a interpretação do modelo. A continuidade desta pesquisa é fundamental para avançar na compreensão e gestão dos recursos de solo, contribuindo para práticas agrícolas sustentáveis e a mitigação das mudanças climáticas. Além disso, futuras pesquisas devem considerar a aplicação dessas técnicas em diferentes contextos geográficos e ambientais para validar a generalização dos modelos desenvolvidos. A integração de novas fontes de dados pode ser um caminho promissor para aprimorar a acurácia e a aplicabilidade das estimativas de SOCS.

Propomos também, para trabalhos futuros, realizar teste de hipótese para verificar se houve uma diferença estatisticamente significativa entre os resultados dos parâmetros RMSE e MAE nos modelos de AM utilizados neste trabalho.

Novos métodos para definir os hiperparâmetros podem ser avaliados em trabalhos futuros, como métodos de inferência bayesiana e de heurísticas populacionais como o Algoritmo Genético Clássico (AG) e o *Biased Random Key Genetic Algorithm* (BRKGA)[13], que é um algoritmo genético que melhora o AG. O objetivo do BRKGA é facilitar a tarefa do operador do AG de lidar com soluções viáveis.

[13] `https://pymoo.org/algorithms/soo/brkga.html`

Referências Bibliográficas

Marcos Bacis Ceddia, Ana Carolina de S. Ferreira, Érika Flávia M. Pinheiro, Michele Duarte de Menezes, and Andréa da Silva Gomes. The use of pedotransfer functions and the estimation of carbon stock in the central amazon region. *Scientia Agricola*, 2016. , 37, 38, 47, 51, 52

Ana Carolina S. Ferreira, Érica Flávia M. Pinheiro, Elias M. Costa, and Marcos Bacis Ceddia. Predicting soil carbon stock in remote areas of the central amazon region using machine learning techniques. *Geoderma Regional - Elsevier*, 2023. 9, 23, 26, 29, 31, 34, 35, 42, 43

Diego B. Haddad, Diego N. Brandao, Laura S. de Assis, Luis Tarrataca, Andrea da S. Gomes, Marcos Bacis Ceddia, Rosane F. Oliveira, and Jurair R. de P. Junior. Brazilian soil bulk density prediction based on a committee of neural regressors. *IEEE - Institute of Electrical and Electronic Engineers*, 2018. 9, 42, 43, 44, 49

Zhishan Ye, Ziheng Sheng, Xiaoyan Liu, Youhua Ma, Ruochen Wang, Shiwei Ding, Zijie Li Liu, and Qiang Wang. Using machine learning algorithms based on gf-6 and google earth engine to predict and map the spatial distribution of soil organic matter content. *Sustainability - MDPI*, 2021. 9, 36

Gabor Szatmari, Laszlo Pasztor, Annamaria Laborczi, Gabor Illes, Zsofia Bakacsi, Dora Zachary, Tibor Filep, Zoltan Szalai, and Gergely Jakab. Countrywide mapping and assessment of organic carbon saturation in the topsoil using machine learning-based pedotransfer function with uncertainty propagation. *Catena - Elsevier*, 2023. 9, 37

Alexandre M.J-C. Wadoux, Budiman Minasny, and Alex B. McBratney. Machine learning for digital soil mapping: applications, challenges and suggested solutions. *Earth-Science Reviews - Elsevier*, 2020a. 9, 14, 21, 31, 32, 33, 35

Sushant Kumar et al. Potential impact of data-centric ai on society. *IEEE Technology and Society Magazine*, 42(3):98–107, 2023a. 9

Luis Japa et al. A population-based hybrid approach for hyperparameter optimization of neural networks. *IEEE Access*, 11:50752–50768, 2023. doi: 10.1109/ACCESS.2023.3277310. 10, 21, 22

Luiz Fernando Carvalho Leite. *Matéria Orgânica do Solo.* Embrapa - ISSN 0104-866X, 2004. 11

Rangel Leandro Romão. *Carbono Orgânico em função do uso do solo.* Universidade Estadual Paulista - UNESP, 2012. 11

Paulo César Teixeira, Guilherme Kangussu Donagemma, Ademir Fontana, and Wenceslau Geraldes Teixeira. *Manual de Métodos de Análise de Solo.* Empresa Brasileira de Pesquisa Agropecuária - Embrapa, 2017. 11

Beatriz Nastaro Boschiero. Matéria orgânica do solo: importância para a fertilidade do solo. `https://agroadvance.com.br/blog-materia-organica-do-solo/`, 2020. [Online; acesso 06-julho-2023]. 12

Aline Marchiori Crespo, Maurício Novaes Souza, and Maria A. Bonfante da Silva. Ciclo de carbono (c) e sistemas agroflorestais na sustentabilidade da produção agrícola: Revisão de literatura. *Instituto Capixaba de Pesquisa, Assistência Técnica e Extensão Rural - INCAPER*, 2023. 12

John Grace. Understanding and managing the global carbon cycle. *Journal of Ecology*, 92(2): 189–202, 2004. 12

David A.N. Ussiri and Rattan Lal. *Introduction to Global Carbon Cycling: An Overview of the Global Carbon Cycle*, pages 61–76. Springer International Publishing, 2017. 12

Claudia Rocha Martins, PA d P Pereira, Wilson Araújo Lopes, and JB d Andrade. Ciclos globais de carbono, nitrogênio e enxofre. *Cadernos temáticos de química nova na escola*, 5:28–41, 2003. 12

Pedro Hofig. *Mapeamento digital de solos e o mapa de solos como ferramenta para classificação de aptidão de uso das terras.* Pós-graduação em Ciência do solo - Universidade Federal do RS, 2014. 13

Mohamed F. Guerri, Cosimo Distante, Paolo Spagnolo, Fares Bougourzi, and Abdelmalik Taleb-Ahmed. Deep learning techniques for hyperspectral image analysis in agriculture: A review. *Elsevier - ISPRS Open Journal of Photogrammetry and Remote Sensing*, 2024. 13

Larissa Anne de S. Lima, Marina R.B. Neumann, Adriana Reatto, and Henrique L. Roig. *Mapeamento de solos do tradicional ao digital.* Embrapa, 2013. 13

Tom M. Mitchell. *Machine Learning.* McGraw-Hill Science, 1997. 14

Katti Faceli, Ana Carolina Lorena, João Gama, and André Carlos P. L. F. Carvalho. *Inteligência Artificial: uma abordagem de Aprendizado de Máquina.* LTC, 2011. 14, 15

Kamal Khosravi Aqdam, Farrokh Asadzadeh, Hamid Reza Momtaz, Naser Miran, and Ehsan Zare. Digital mapping of soil erodibility factor in northwestern iran using machine learning models. *Environ Monit Assess*, 2022. 14, 17, 35

Rafael Izbicki and Tiago Mendonça dos Santos. *Aprendizado de máquina: uma abordagem estatística*. Câmara Brasileira do Livro - SP, 2020. 14

International Business Machines Corporation IBM. O que é Aprendizado Não Supervisionado? `https://www.ibm.com/br-pt/topics/unsupervised-learning`, 2024. [Online; acesso 11-dezembro-2024]. 14

Didática Tech Inteligência Artificial Data Science. Problemas de Classificação e Regressão. `https://didatica.tech/problemas-de-classificacao-e-regressao/`, 2024. [Online; acesso 14-novembro-2024]. 15

Alexandre M. J.-C. Wadoux, Nicolas P. A. Saby, and Manuel P. Martin. Shapley values reveal the drivers of soil organic carbon stock prediction. *Soil*, 2023. 16, 20, 31, 32

Francis B.T. Silatsa, Martin Yemefack, Fritz O. Tabi, Gerard B.M. Heuvelink, and Johan G.B. Leenaars. Assessing countrywide soil organic carbon stock using hybrid machine learning modelling and legacy soil data in cameroon. *Geoderma - Elsevier*, 2020. 17, 33

Jessica BOOKDOWN Kubrusly. intr machine learning. `https://bookdown.org/jessicakubrusly/intr-machine-learning-i/_book/cap-boosting.html`, 2024. [Online; acesso 06-dezembro-2024]. 17

Boqiang Xie, Jianli Ding, Xiangyu Ge, Xiaohang Li, Lijing Han, and Zheng Wang. Estimation of soil organic carbon content in the ebinur lake wetland, xinjiang, china, based on multisource remote sensing data and ensemble learning algorithms. *Sensors - MDPI*, 2022. 17, 34

Hamid Mahmoudzadeh, Hamid Reza Matinfar, Ruhollah Taghizadeh-Mehrjardi, and Ruth Kerry. Spatial prediction of soil organic carbon using machine learning techniques in western iran. *Geoderma Regional - Elsevier*, 2020. 17, 32, 34

Bharani Kumar 360DigiTMG Depuru. Comparison GridSearchCV and RandomSearch CV. `https://360digitmg.com/blog/comparison-gridsearchcv-and-randomsearchcv`, 2023. [Online; acesso 14-novembro-2024]. 19

Davide Anguita, Luca Ghelardoni, Alessandro Ghio, Luca Oneto, Sandro Ridella, et al. The'k'in k-fold cross validation. In *ESANN*, volume 102, pages 441–446, 2012. 19

Moore David S., William I. Notz, and Michael A. Fligner. *The Basic Practice of Statistics - Seventh Edition*. W.H. Freeman Company, 2007. 20

Nur'aina Daud, Nurulhuda Noordin, and Anitawati Lokman. Obesity predictor identification: Comparison of correlation based feature selection method and wrapper method on nutrition dataset. *Journal of Advanced Research in Applied Sciences and Engineering Technology*, 2025. 21

N. Shahidian, P. Abreu, D. Santos, and A. Barbosa-Povoa. Short-term forecasting of emergency medical services demand exploring machine learning. *Journal of Advanced Research in Applied Sciences and Engineering Technology*, 2024. 21

Rizgar R. Zebari, Adnan Mohsin Abdulazeez, Diyar Qader Zeebaree, Dilovan Asaad Zebari, and Jwan Najeeb Saeed. A comprehensive review of dimensionality reduction techniques for feature selection and feature extraction. *The Journal of Applied Science and Technology Trends (JASTT)*, 2020. 21

Sole Galli. Feature Selection with Wrapper Methods in Python. `https://www.blog.trainindata.com/feature-selection-with-wrapper-methods/`, 2023. [Online; acesso 10-novembro-2024]. 21

Kevin Jamieson and Ameet Talwalkar. Non-stochastic best arm identification and hyperparameter optimization. In *AISTATS 2016 - Proceedings*, volume 51 of *Proceedings of Machine Learning Research*, pages 240–248. PMLR, 2016. URL `https://proceedings.mlr.press/v51/jamieson16.html`. 21

Manoj Kumar et al. Parallel architecture and hyperparameter search via successive halving and classification. *arXiv*, 2018. 21

Tanay Agrawal. *Hyperparameter optimization in machine learning: make your machine learning and deep learning models more efficient*. Springer, 2021. 21

Daniel S. Soper. Hyperparameter optimization using successive halving with greedy cross validation. *Algorithms*, 16(1), 2023. ISSN 1999-4893. doi: 10.3390/a16010017. URL `https://www.mdpi.com/1999-4893/16/1/17`. 21, 22

Danielle Pinna, Rodrigo Toso, Gustavo Semaan, Fernando de Sá, Ana I Pereira, Ângela Ferreira, Jorge Soares, and Diego Brandão. Fault classification of wind turbine: A comparison of hyperparameter optimization methods. In *International Conference on Optimization, Learning Algorithms and Applications*, pages 229–243. Springer, 2023. 21

Liam Li, Kevin Jamieson, Afshin Rostamizadeh, Ekaterina Gonina, Jonathan Ben-Tzur, Moritz Hardt, Benjamin Recht, and Ameet Talwalkar. A system for massively parallel hyperparameter tuning. *Proceedings of Machine Learning and Systems*, 2:230–246, 2020. 22

Liam Li. Massively parallel hyperparameter optimization. *Carnegie Mellon University - CMU*, 2018. 22

Jingru Song et al. Estimation of soil organic carbon content in coastal wetlands with measured vis-nir spectroscopy using optimized support vector machines and random forests. *Remote Sensing - MDPI*, 2022. 22

Tarun Luthra and Rahul Janghu. Successive Halving Search. `https://www.scaler.com/topics/machine-learning/successive-halving-search/`, 2023. [Online; acesso 12-dezembro-2024]. 22

Tobias Rentschler, Ulrike Werban, Mario Ahner, Thorsten Behrens, Philipp Gries, Thomas Scholten, Sandra Teuber, and Karsten Schmidt. 3d mapping of soil organic carbon content and soil moisture with multiple geophysical sensors and machine learning. *Vadose Zone Journal*, 2020. 26, 34

Stephen A Book and Philip H Young. The trouble with r2. *Journal of Parametrics*, 25(1):87–114, 2006. 27

Ariel M Aloe, Betsy Jane Becker, and Therese D Pigott. An alternative to r2 for assessing linear models of effect size. *Research Synthesis Methods*, 1(3-4):272–283, 2010. 27

Joel Tellinghuisen and Carl H Bolster. Using r2 to compare least-squares fit models: When it must fail. *Chemometrics and Intelligent Laboratory Systems*, 105(2):220–222, 2011. 27

Shinichi Nakagawa and Holger Schielzeth. A general and simple method for obtaining r2 from generalized linear mixed-effects models. *Methods in ecology and evolution*, 4(2):133–142, 2013. 27

He-shuang Wan, Wei-chun Zhang, Ming-hou Zhou, Wei Wu, and Hong-bin Liu. Soil total and organic carbon mapping and uncertainty analysis using machine learning techniques. *Ecological Indicators - Elsevier*, 2022. 29, 33

C. Wade Ross, Sabine Grunwald, Timothy A. Martin, Rosvel Bracho, Allan R. Bacon, Colby W. Brungard, Jason G. Vogel, Daniel Markewitz, Eric J. Jokela, and Xiong Xiong. Accounting for two-billion tons of stabilized soil carbon. *Elsevier*, 2020. 31, 35

Ndiye M. Kebonye, Príncipe C. Agyeman, and James K.M. Biney. Optimized modelling of countrywide soil organic carbon levels via an interpretable decision tree. *Elsevier - Smart Agricultural Technology*, 2023. 32, 37

Stephan van der Westhuizen, Gerard B.M. Heuvelink, and David P. Hofmeyr. Multivariate random forest for digital soil mapping. *Elsevier - Geoderma*, 2023. 32, 36

Jie Chen, Huan Zhang, Man Man Fan, Furong Chen, and Chao e Gao. Machine learning based prediction and key factor identification of the organic carbon in riverine floodplain soils with intensive agricultural practices. *Journal of Soils and Sediments*, 2021. 32, 33

Larissa Macedo dos Santos, Marcos S. Colla, Jéssica B. Carra, Marciéli Fabris, and Vanderlei A. de Lima. Classification and total carbon determination of the soils using rgb digital images combined with machine learning. *Communications In Soil Science and Plant Analysis*, 2022. 32, 35

Jiayi Yang, fã Junjian, Zefan Lan, Xingmin Mu, Yiping Wu, Zhongbao Xin, Puqiong Miping, and Guangju Zhao. Improved surface soil organic carbon mapping of soilgrids250m using sentinel-2 spectral images in the qinghai–tibetan plateau. *Remote Sensing - MDPI*, 2022a. 32, 34

Mostafa Emadi, Ruhollah Taghizadeh-Mehrjardi, Ali Cherati, Majid Danesh, Amir Mosavi, and Thomas Scholten. Predicting and mapping of soil organic carbon using machine learning algorithms in northern iran. *Remote Sensing - MDPI*, 2020. 32, 35

Ruhollah Taghizadeh-Mehrjardi, Karsten Schmidt, Alireza Amirian-Chakan, Tobias Rentschler, Mojtaba Zeraatpisheh, Fereydoon Sarmadian, Roozbeh Valavi, Naser Davatgar, Thorsten Behrens, and Thomas Scholten. Improving the spatial prediction of soil organic carbon content in two contrasting climatic regions by stacking machine learning models and rescanning covariate space. *Remote Sensing - MDPI*, 2020. 32, 35

Mingxing Xu, Xianyao Chu, Yesi Fu, Changjiang Wang, and Shaohua Wu. Improving the accuracy of soil organic carbon content prediction based on visible and near-infrared spectroscopy and machine learning. *Environmental Earth Sciences*, 2021. 32, 36

Ali Sakhaee, Anika Gebauer, Mareike Ließ, and Axel Don. Spatial prediction of organic carbon in german agricultural topsoil using machine learning algorithms. *Copernicus Publications*, 2022. 33, 34

Xiaoyu Huang, Xuemei Wang, Kawuqiati Baishan, and An Baisong. Hyperspectral estimation of soil organic carbon content based on continuous wavelet transform and successive projection algorithm in arid area of xinjiang, china. *Sustainability - MDPI*, 2023. 33, 36

José Padarian, Budiman Minasny, and Alex B. McBratney. Machine learning and soil sciences: a review aided by machine learning tools. *Soil*, 2020. 33

Omosalewa Odebiri, John Odindi, and Onisimo Mutanga. Basic and deep learning models in remote sensing of soil organic carbon estimation: A brief review. *International Journal of Applied Earth Observations and Geoinformation - Elsevier*, 2021. 33

Rajneesh Sharma, Deepak R. Mishra, Mateus R. Levi, and Lori A. Sutter. Remote sensing of surface and subsurface soil organic carbon in tidal wetlands: A review and ideas for future research. *Remote Sensing - MDPI*, 2022. 33

T. Li, A. Xia, T.I. McLaren, R. Pandey, Z. Xu, H. Liu, S. Manning, O. Madgett, S. Duncan, P. Rasmussen, and et al. Preliminary results in innovative solutions for soil carbon estimation integrating remote sensing, machine learning and proximal sensing sensing. *Remote Sensing - MDPI*, 2023. 33

T. Li, L. Cui, M. Kuhnert, T. I. McLaren, R. Pandey, H. Liu, W. Wang, Z. Xu, A. Xia, R. C. Dalal, and Y. P. Dang. A comprehensive review of soil organic carbon estimates integrating remote sensing and machine learning technologies. *Journal of Soils and Sediments*, 2024a. 33

J. Qadri and F. Ceccato. Exploring the interaction between landslides and carbon stocks in italy. *Sustainability - MDPI*, 2024. 33

D. Radocaj, D. Jug, I. Jug, and M. A. Jurišic. A comprehensive evaluation of machine learning algorithms for digital soil organic carbon mapping on a national scale. *Applied Science - MDPI*, 2024. 33, 34, 35

T. Li, L. Cui, Y. Wu, T. I. McLaren, A. Xia, R. Pandey, H. Liu, W. Wang, Z. Xu, X. Song, and et al. Soil organic carbon estimation via remote sensing and machine learning techniques global topic modeling and research trend exploration. *Remote Sensing - MDPI*, 2024b. 33

Kingsley John, Isong Abraham, Ndiye Michael Kebonye, Esther Okon Ayito, Prince Chapman Agyeman, and Sunday Marcus Afu. Using machine learning algorithms to estimate soil organic carbon variability with environmental variables and soil nutrient indicators in an alluvial soil. *Land - MDPI*, 2020. 33

S. Roohollah Mousavi, Fereydoon Sarmadian, Mahmoud Omid, and Patrick Bogaert. Three-dimensional mapping of soil organic carbon using soil and environmental covariates in an arid and semi-arid region of iran. *Journal of the International Measurement Confederation - Elsevier*, 2022. 33

F. Hateffard, G. Szatmári, and T. J. Novák. Applicability of machine learning models for predicting soil organic carbon content and bulk density under different soil conditions. *Soil Science Annual*, 2023. 33, 35

P. Pellikka, M. Luotamo, N. Sädekoski, R. Hietanen, J. Zepp, I. Vuorinne, M. Räsänen, J. Heiskanen, M. Siljander, K. Karhu, and A. Klami. Tropical altitudinal gradient soil organic carbon and nitrogen estimation using specim iq portable imaging spectrometer. *Science of the Total Environment*, 2023. 33

L. Wang, G. Abramowitz, Y-P. Wang, A. Pitman, and R. A. Viscarra Rossel. An ensemble estimate of australian soil organic carbon using machine learning and process-based modelling. *Soil Open Access*, 2024a. 33

J. Ou, Z. Wu, Q. Yan, X. Feng, and Z. Zao. Improving soil organic carbon mapping in farmlands using machine learning models and complex cropping system information. *Environmental Sciences Europe*, 2024. 33, 34

K.M.d. Oliveira, J.V.F. Gonçalves, R.H. Furlanetto, C.A.d. Oliveira, W.A. Mendonça, D.d.F.d.S. Haubert, L.G.T. Crusiol, R. Falcioni, R.B.d. Oliveira, A.S. Reis, and et al. Predicting particle size and soil organic carbon of soil profiles using vis-nir-swir hyperspectral imaging and machine learning models. *Remote Sensing - MDPI*, 2024. 33

Waldir de C. Junior, Braz C. Filho, Cesar da S. Chagas, Silvio B Bhering, Nilson R. Pereira, and Helena Saraiva K. Pinheiro. Multiple linear regression and random forest model to estimate soil bulk density in mountainous regions. *Pesquisa Agropecuária Brasileira*, 2016. 33

Zong Wanga, Zhengping Du, Xiaoyan Li, Zhengyi Bao, Na Zhao, and Tianxiang Yue. Incorporation of high accuracy surface modeling into machine learning to improve soil organic matter mapping. *Elsevier*, 2021. 33

Umakant Mishra, Sagar Gautam, William J. Riley, and Forrest M. Hoffman. Ensemble machine learning approach improves predicted spatial variation os surface soil organic carbon stocks in data-limited northern circumpolar region. *Frontiers in Big Data*, 2020. 33

Biswabara Sahu, Amlan Kumar Ghosh, and Seema. Deterministic and geostatistical models for predicting soil organic carbon in a 60 ha farm on inceptisol in varanasi, india. *Geoderma Regional - Elsevier*, 2021. 33

Hamid Reza Matinfar, Ziba Maghsodi, Sayed Roholla Mousavi, and Asghar Rahmani. Evaluation and prediction of topsoil organic carbon using machine learning and hybrid models at a field-scale. *Catena - Elsevier*, 2021. 33

Iqra Farooq, Shabir A. Bangroo, Owais Bashir, Tajamul I. Shah, Ajaz A. Malik, Asif M. Iqbal, Syed S. Mahdi, Owais A. Wani, Nageena Nazir, and Asim Biswas. Comparison of random forest and kriging models for soil organic carbon mapping in the himalayan region of kashmir. *Communications In Soil Science and Plant Analysis*, 2022. 33

Changda Zhu, Yuchen Wei, Fubin Zhu, Wenhao Lu, Zihan Fang, and Zhaofu Li. Digital mapping of soil organic carbon based on machine learning and regression kriging. *Sensors - MDPI*, 2022. 33

Y. Zhang, Y. Wang, Y. Bai, R. Zhang, X. Liu, and X. Ma. Prediction of spatial distribution of soil organic carbon in helan farmland based on different prediction models. *Land - MDPI*, 2023a. 33, 35, 36

Gábor Szatmári, László Pásztor, and Gerard B.M. Heuvelink. Estimating soil organic carbon stock change at multiple scales using machine learning and multivariate geostatistics. *European Journal of Soil Science*, 2021. 33

Fuat Kaya, Ali Keshavarzi, Rosa Franviglia, Gordana Kaplan, Levent Başayiğit, and Mert Dedeoğlu. Assessing machine learning-based prediction under different agricultural practices for digital mapping of soil organic carbon and available phosphorus. *Agriculture - MDPI*, 2022. 33

Gaurav Mishra, Magboul M. Sulieman, Foat Kaya, Rosa Franviglia, Ali Keshavarzi, Esmael Bakhshandeh, Macoumba Loum, Abhishek Jangir, Ibrahim Ahmed, Abdelmagid Elmobarak, AdbElrahim Abdelmagid, and Deepa Rawat. Machine learning for cation exchange capacity prediction in different land uses. *SSRN*, 2022a. 33

Yi Xiao, Jie Xue, Xianglin Zhang, Nan Wang, Yongsheng Hong, Yefeng Jiang, Yin Zhou, Hongfen Teng, Bifeng Hu, Emanuele Lugato, Anne C. Richer-de Forges, Dominique V, Zhou Shi, and Songchao Chen. Improving pedotransfer functions for predicting soil mineral associated organic carbon by ensemble machine learning. *Geoderma - Elsevier*, 2022. 33

M. Meliho, M. Boulmane, A. Khattabi, C.E. Dansou, C.A. Orlando, N. Mhammdi, and K.D. Noumonvi. Spatial prediction of soil organic carbon stock in the moroccan high atlas using machine learning. *Remote Sensing - MDPI*, 2023. 34

S. Nozari, M. R. Pahlavan-Rad, C. Brungard, B. Heung, and L. Borůvka. Digital soil mapping using machine learning-based methods to predict soil organic carbon in two different districts in the czech republic. *Soil and Water Research*, 2024. 34

P. C. Moharana, B. Yadav, L. C. Malav, S. Kumar, R. L. Meena, M. Nogiya, H. Biswas, and N. G.

Patil. Regional prediction of soil organic carbon dynamics for intensive farmland in the hot arid climate of india using the machine learning model. *Environmental Earth Sciences*, 2024. 34

W. Li, Z. Yang, J. Jiang, and G. Sun. Spatial variation and stock estimation of soil organic carbon in cropland in the black soil region of northeast china. *Agronomy - MDPI*, 2024c. 34

Mareike Lieb, Johannes Schmidt, and Bruno Glaser. Improving the spatial prediction of soil organic carbon stocks in a complex tropical mountain landscape by methodological specifications in machine learning approaches. *PLoS ONE*, 2016. 34

Camile Sothe, Alemu Gonsamo, Joyce Arabian, and James Snider. Large scale mapping of soil organic carbon concentration with 3d machine learning and satellite observations. *Geoderma - Elsevier*, 2022. 34

A. Kumar, P.C. Moharana, R.K. Jena, S.K. Malyan, G.K. Sharma, R.K. Fagodiya, A.A. Shabnam, D.K. Jigyasu, K.M.V. Kumari, and S.G. Doss. Digital mapping of soil organic carbon using machine learning algorithms in the upper brahmaputra valley of northeastern india. *Land - MDPI*, 2023b. 34

Maryam G. Goydaragh, Ruhollah Taghizadeh-Mehrjardi, Ali A. Jafarzadeh, John Triantafilis, and Marcos Lado. Using environmental variables and fourier transform infrared spectroscopy to predict soil organic carbon. *Catena - Elsevier*, 2021. 34

Xianglin Zhang, Jie Xue, Songchao Chen, Nan Wang, Zhou Cusido, Shi, Yuanfang Huang, and Zhiqing Zhuo. Digital mapping of soil organic carbon with machine learning in dryland of northeast and north plain china. *Remote Sensing - MDPI*, 2022. 34

Tessa S. van der Voort, Sven Verweij, Yuki Fujita, and Gerard H. Ros. Enabling soil carbon farming: presentation of a robust, afordable, and scalable method for soil carbon stock assessment. *Elsevier - Smart Agricultural Technology*, 2023. 34

Ton Broeg, Michael Blaschek, Steffen Seitz, Ruhollah Taghizadeh-Mehrjardi, Simone Zepp, and Thomas Scholten. Transferability of covariates to predict soil organic carbon in cropland soils. *Remote Sensing - MDPI*, 2023a. 34

E.P.d. Santos, M.C. Moreira, E.I. Fernandes-Filho, J.A.M. Demattê, E.A. Dionizio, D.D.d. Silva, R.R.P. Cruz, J.M. Moura-Bueno, U.J.d. Santos, and M.H. Costa. Sentinel-1 imagery used for estimation of soil organic carbon by dual-polarization sar vegetation indices. *Remote Sensing - MDPI*, 2023. 34

Q. Chen, W. Zhou, and W. Shi. Estimation of soil organic carbon density on the qinghai–tibet plateau using a machine learning model driven by multisource remote sensing. *Remote Sensing - MDPI*, 2024. 34, 36

M. Golkar Amoli, M. Hasanlou, R. Taghizadeh Mehrjardi, and F. Samadzadegan. Exploring the potential of prisma satellite hyperspectral image for estimating soil organic carbon in marvdasht region, southern iran. *Remote Sensing - MDPI*, 2024. 34

Xia Liu, Tao Zhou, Peijun Shi, Yajie Zhang, Hui Luo, Peixin Yu, Yixin Xu, Peifang Zhou, and Jingzhou Zhang. Uncertainties of soil organic carbon stock estimation caused by paleoclimate and human footprint on the qinghai plateau. *Carbon Balance and Management*, 2022. 34

Sushil Lamichhane, Kabindra Adhikari, and Lalit Kumar. National soil organic carbon map of agricultural lands in nepal. *Geoderma Regional - Elsevier*, 2022. 34

P. Fu, C. Clanton, K. M. Demuth, V. Goodman, L. Griffith, M. Khim-Young, J. Maddalena, K. LaMarca, L.A. Wright, D.W. Schurman, and et al. Accurate quantification of 0–30 cm soil organic carbon in croplands over the continental united states using machine learning. *Remote Sensing - MDPI*, 2024. 34, 35

H. Gao, J. Gong, J. Yang, J. Liu, and T. Ye. Effects of land use cover changes on soil organic carbon stocks in qinghai tibet plateau: A comparative analysis of different ecological functional areas based on machine learning methods and soil carbon pool data. *Research Square*, 2024. 34

O. D. Adeniyi, A. Brenning, and M. Maerker. Spatial prediction of soil organic carbon combining machine learning with residual kriging in an agricultural lowland area (lombardy region, italy). *Geoderma*, 2024. 34

Cai J.T. Ladd, Craig Smeaton, Martin W. Skov, and William E.N. Austin. Best practice for upscaling soil organic carbon stocks in salt marshes. *Geoderma - Elsevier*, 2022. 34

Mojtaba Zeraatpisheh, Shamsollah Ayoubi, Zahra Mirbagheri, Mohammad R. Mosaddeghi, and Ming Xu. Spatial prediction of soil aggregate stability and soil organic carbon in aggregate fractions using machine learning algorithms and environmental variables. *Geoderma Regional - Elsevier*, 2021. 34

B. Barragán-Pazmiño, A. O. Echeverría, M. E. Guadalupe, and T. Toulkeridis. Integrating machine learning models for enhanced soil organic carbon estimation a multi-model fusion approach. *ASPG*, 2024. 34

Reshmi Sarkar and Anil Somenahally. Machine learning soil-environmental impacts on agroecosystems for relating microbial biomass to soil carbon sequestration. *Elsevier - Smart Agricultural Technology*, 2023. 34

E. Xie, J. Chen, Y. Peng, G. Yan, and Y. Zhao. Integration of machine learning and process-based model outputs via ensemble kalman filter enhanced space–time modelling of soil organic carbon in a highly human impacted area. *Soil Use and Management*, 2024. 34

P. Abbaszad, F. Asadzadeh, S. Rezapour, K. K. Aqdam, and F. Shabani. Evaluation of landsat 8 and sentinel-2 vegetation indices to predict soil organic carbon using machine learning models. *Modeling Earth Systems and Environment*, 2024. 35

M. Carbajal, D. A. Ramırez, C. Turin, S. M. Schaeffer, J. Konkel, J. Ninanya, J. Rinza, Mendiburu. F., P. Zorogastua, L. Villaorduña, and R. Quiroz. From rangelands to cropland, land-use change and its impact on soil organic carbon variables in a peruvian andean highlands a machine learning modeling approach. *Ecosystems*, 2024. 35

Chelene C. Hanes, Mike Woton, Douglas G. Woolford, David L. Martell, and Mike Flannigan. Mapping organic layer thickness and fuel load of the boreal forest in alberta, canada. *Geoderma - Elsevier*, 2022. 35

Alexandre MJ-C. Wadoux, Alessandro Samuel-Rosa, Laura Poggio, and Vera Leatitia Mulder. A note on knowledge discovery and machine learning in digital soil mapping. *European Journal of Soil Science*, 2020b. 35

Melissa S. Schwab, Hannah Gies, Chantal V. Freymond, Maarten Lupker, Negar Haghipour, and Timothy Ian Eglinton. Environmental and hydrologic controls on sediment and organic carbon export from a subalpine catchment: insights from a time series. *Biogeosciences*, 2022. 35

Umakant Mishra, Forrest M. Hoffman, Kyongmin Yeo, Corey Hudson, Kabindra Adhikari, Sagar Gautam, and William J. Riley. Empirical relation ships between environmental factors and soil organic carbon produce comparable prediction accuracy to machine learning. *Soil Science Society of America Journal*, 2022b. 35

Abiot Molla, Weiwei Zhang, Shudi Zuo, Yin Ren, and Jigang Han. A machine learning and geostatistical hybrid method to improve spatial prediction accuracy of soil potentially toxic elements. *Research Square Company*, 2022. 35

Sedigheh Maleki, Alireza Karimi, Amin Mousavi, Ruth Kerry, and Ruhollah Taghizadeh-Mehrjardi. Delineation of soil management zone maps at the regional scale using machine learning. *Agronomy - MDPI*, 2023. 35

Chun Ki Chan, Carla A. Gomez, Anish Kothikar, and P.M. Baiz-Villafranca. Satellite-based carbon estimation in scotland: Agb and soc. *Land - MDPI*, 2023. 36

Jannis Heil, Christoph Jörges, and Britta Stumpe. Fine-scale mapping of soil organic matter in agricultural soils using uavs and machine learning. *Remote Sensing - MDPI*, 2022. 36

Masoud Zolfaghari Nia, Mostafa Moradi, Gholamhosein Moradi, and Ruhollah Taghizadeh-Mehrjardi. Machine learning models for prediction of soil properties in the riparian forests. *Land - MDPI*, 2022. 36

Odunayo David Adeniyi, Alexander Brenning, Alice Bernini, Stefano Brenna, and Michael Maerker. Digital mapping of soil properties using ensemble machine learning approaches in an agricultural lowland area of lombardy, italy. *Land - MDPI*, 2023. 36

Mareike Lieb. Modeling the agricultural soil landscape of germany—a data science approach involving spatially allocated functional soil process units. *Agriculture - MDPI*, 2022. 36

Markus Möller, Simone Zepp, Martin Wiesmeier, Heike Gerighausen, and Uta Heiden. Scale-specific prediction of topsoil organic carbon contents using terrain attributes and scmap soil reflectance composites. *Remote Sensing - MDPI*, 2022. 36

Paulina B. Rameurez, Samantha Mosier, Francisco Calderón, and Francesca V. Using mid-infrared spectroscopy to optimize throughput and costs of soil organic carbon and nitrogen estimates: An assessment in grassland soils. *Environments - MDPI*, 2022. 36

Bertin Takoutsinga and Gerard B. M. Heuvelink. Comparing the prediction performance, uncertainty quantification and extrapolation potential of regression kriging and random forest while accounting for soil measurement errors. *Geoderma - Elsevier*, 2022. 36

Kpade O. L. Hounkpatin, Johan Stendahl, Mattias Lundblad, and Erik Karltun. Predicting the spatial distribution of soil organic carbon stock in swedish forests using a group of covariates and site-specific data. *Copernicus Publications*, 2021. 36

Y. Zhang, C. Kou, M. Liu, W. Man, F. Li, C. Lu, J. Song, T. Song, Q. Zhang, X. Li, and et al. Estimation of coastal wetland soil organic carbon content in western bohai bay using remote sensing, climate, and topographic data. *Remote Sensing - MDPI*, 2023b. 36

T. Broeg, M. Blaschek, S. Seitz, R. Taghizadeh-Mehrjardi, S. Zepp, and T. Scholten. Transferability of covariates to predict soil organic carbon in cropland soils. *Remote Sensing - MDPI*, 2023b. 36

W. Yu, W. Zhou, T. Wang, J. Xiao, Y. Peng, H. Li, and Y. Li. Significant improvement in soil organic carbon estimation using data-driven machine learning based on habitat patches. *Remote Sensing - MDPI*, 2024. 36

C. Nwaogu, B. E. Diagi, C. V. Ekweogu, A. S. Ajeyomi, C. C. Ejiogu, E. I. Emereibeole, P. S. U. Eneche, O. J. Okeke, D. O. Edokpa, E. Chike, and et al. Soil organic carbon stocks as driven by land use in mato grosso state the brazilian cerrado agricultural frontier. *Discover Sustainability*, 2024. 36

A. Chinilin and I. Y. Savin. Combining machine learning and environmental covariates for mapping of organic carbon in soils of russia. *The Egyptian Journal of Remote Sensing and Space Sciences*, 2023. 36

S. Agaba, C. Ferré, M. Musetti, and R. Comolli. Mapping soil organic carbon stock and uncertainties in an alpine valley (northern italy) using machine learning models. *Land - MDPI*, 2024. 36

Z. Wang, J. Kumar, S. R. Weintraub-Leff, K. Todd-Brown, U. Mishra, and D. Sihi. Upscaling soil organic carbon measurements at the continental scale using multivariate clustering analysis and machine learning. *JGR Biogeosciences*, 2024b. 36

Nyenshu S. Rengma, K. Justin George, and Suresh Kumar. Remote sensing based digital mapping of soil properties in a himalayan watershed employing machine learning technique. *Indian Institute of Remote Sensing*, 2020. 36

G. Ravaioli, T. Domingos, and Ricardo F. M. Teixeira. Data-driven agent-based modelling of incentives for carbon sequestration the case of sown biodiverse pastures in portugal. *Journal of Environmental Management*, 2023. 36

M. Ding, X. Li, and Z. Jin. Digital mapping of soil organic carbon using uav images and soil properties in a thermo-erosion gully on the tibetan plateau. *Remote Sensing - MDPI*, 2023. 36

X. Xiao, S. Ma, Q. He, J. Liu, W. Sun, Y. Lin, and R. Yi. Environmental variables improve the accuracy of remote sensing estimation of soil organic carbon content. *Scientific Reports*, 2024. 37

L. Mango, N. Narissara, and J. Som-ard. Estimating soil organic carbon using sentinel-2 data under zero tillage agriculture a machine learning approach. *Research Square*, 2024. 37

T. Zhang, Ye Li, and M. Wang. Prediction of soil organic carbon and total nitrogen affected by mine using vis–nir spectroscopy coupled with machine learning algorithms in calcareous soils. *Scientific Reports*, 2024. 37

Sandeep Kumara, Rattan Lala, and Desheng Liu. A geographically weighted regression kriging approach for mapping soil organic carbon stock. *Geoderma - Elsevier*, 2012. 37

Tao Liu, Huan Zhang, and Tiezhu Shi. Modeling and predictive mapping of soil organic carbon density in a small-scale area using geographically weighted regression kriging approach. *Sustainability - MDPI*, 2020. 37

Haithem Bahri, Damien Raclot, Meriem Barbouchi, Philippe Lagacherie, and Mohamed Annabi. Mapping soil organic carbon stocks in tunisian topsoils. *Geoderma Regional - Elsevier*, 2022. 37

X. Zhang, J. Xue, S. Chen, N. Wang, T. Xie, Y. Xiao, X. Chen, Z. Shi, Y. Huang, and Z. Zhuo. Fine resolution mapping of soil organic carbon in croplands with feature selection and machine learning in northeast plain china. *Remote Sensing - MDPI*, 2023c. 37

Anna Brook, Seham Hamzi, Dar Roberts, Charles Ichoku, Nurit Shtober-Zisu, and Lea Wittenberg. Total carbon content assessed by uas near-infrared imagery as a new fire severity metric. *Remote Sensing - MDPI*, 2022. 37

David Beltrán-Marcosa, Susana Suárez-Seoane, José M. Fernández-Guisuraga, Víctor Fernández-García, and Leonor Calvo. Relevance of uav and sentinel-2 data fusion for estimating topsoil organic carbon after forest fire. *Elsevier - Geoderma*, 2023. 37

K. Adhikari, U. Mishra, P.R. Owens, Libohova., S.A. Testamentos, W.J. Riley, F.M. Hoffman, and D.R. Smith. Importance and strength of environmental controllers of soil organic carbon changes with scale. *Geoderma - Elsevier*, 2020. 37

Said Nawar, Muhammad Abdul Munnaf, and Abdul Mounem Mouazen. Machine learning based on-line prediction of soil organic carbon after removal of soil moisture effect. *Remote Sensing - MDPI*, 2020. 37

Kingsley John, Ndiye M. Kebonye, Príncipe C. Agyeman, and Samuel K. Ahado. Comparison of cubist models for soil organic carbon prediction via portable xrf measured data. *European Journal of Plant Pathology*, 2021. 37

Tiago G. Morais, Camila Tufik, Ana E. Rato, Nuno R. Rodrigues, Ivo Gama, Marjan Jongen, João Serrano, David Fangueiro, Tiago Domingos, and Ricardo FM Teixeira. Estimating soil organic carbon of sown biodiverse permanent pastures in portugal using near infrared spectral data and artificial neural networks. *Geoderma - Elsevier*, 2021. 37

Eva Kanari, Lauric Cécillon, and Pierre Barré. A robust initialization method for accurate soil organic carbon simulations. *Biogeosciences*, 2022. 37

Shurong Yang, Danrui Sheng, Jan Adamowski, Yifan Gong, Jian Zhang, and Jianjun Cao. Effect of land use change on soil carbon storage over the last 40 years in the shi yang river basin, china. *Land - MDPI*, 2022b. 37

Sidarto S. Paulo, Brandon Heung, and Derek H. Lynch. Modeling of total and active organic carbon dynamics in agricultural soil using digital soil mapping: a case study from central nova scotia. *Canadian Journal of Soil Science*, 2022. 37

Min-Jee Kim, Hye-In Lee, Jae H. Choi, Kyoung J. Lim, and Changyeun Mo. Development of a soil organic matter content prediction model based on supervised learning using vis-nir/swir spectroscopy. *Sensors - MDPI*, 2022. 37

Samad Emamgholizadeh, Ahmad Bazoobandi, Babak Mohammadi, Hadi Ghorbani, and Mohammad A. Sadeghi. Prediction of soil cation exchange capacity using enhanced machine learning approaches in the southern region of the caspian sea. *Ain Shams Engineering Journal*, 2023. 38

A. Lopatka, G. Siebielec, R. Kaczynski, and T. Stuczynski. Analysis of soil carbon stock dynamics by machine learning polish case study. *Land - MDPI*, 2023. 38

M. Gouda, M. Abu-hashim, A. Nassrallah, M. N. Khalil, E. Hendawy, F. F. Benhasher, M. S. Shokr, M. A. Elshewy, and Mohamed. E. S. Integration of remote sensing and artificial neural networks for prediction of soil organic carbon in arid zones. *Frontiers Environmental Science*, 2024. 38

J. Lei, C. Zeng, X. Wang, C. Ma, T. Zhou, B. Laffitte, Ke Luo, Z. Yang, and X. Tang. Prediction of soil organic carbon stock combining sentinel-1 and sentinel-2 images in the zoige plateau, the northeastern qinghai-tibet plateau. *Ecological Processes*, 2024. 38

Marcos Bacis Ceddia, André Luis O. Villela, Érika Flávia M. Pinheiro, and Ole Wendroth. Spatial variability of soil carbon stock in the urucu river basin, central amazon-brazil. *Science of the Total Environment*, 2015. 38, 43

Cenk Temizel, Celal Hakan Canbaz, Onder Saracoglu, Dike Putra, Ali Baser, Tomi Erfando, Shanker Krishna, and Luigi Saputelli. Production forecasting in shale reservoirs using lstm method in deep learning. *Unconventional Resources Technology Conference*, 2020. 40

D. B. Haddad, D. N. Brandao, Tayana Moreira, M. B. Ceddia, E. F. M. Pinheiro, and R. F. Oliveira. A first approach using neural network to estimating soil bulk density of urucu basin in central amazon-brazil. *IEEE - Institute of Electrical and Electronic Engineers*, 2017. 42, 44, 49

Ana Alves. *Estatística Aplicada: Análise de Dados*. Aprender Estatística Fácil, 2024. 50, 51

Thibaut Vidal, Toni Pacheco, and Maximilian Schiffer. Born-again tree ensembles. *Cornell University*, 2020. 54, 55