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

Alexandre Pardelinha[1], **Marcos Bacis Ceddia**[2], **Roberto Gervasio**[1], **Kele Belloze** [1], **Carolina de L. Aguilar**[1], **Laura Assis**[1], **Diego Brandão**[1]

[1]Programa de Pós-graduação em Ciência da Computação
Centro Federal de Educação Tecnológica Celso Suckow da Fonseca – CEFET/RJ
Rio de Janeiro – RJ – Brasil

[2]Instituto de Agronomia - UFRRJ, Seropédica

diego.brandao@cefet-rj.br

Abstract. *Soils represent the most significant stock of organic carbon (SOC) 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. 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.*

Resumo. *Os solos representam o mais significativo estoque de carbono orgânico (SOC) nos ecossistemas terrestres, sublinhando a importância crítica de estimar com precisão 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. 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).*

1. Introdução

O solo desempenha um papel crucial na sustentação da sociedade e dos ecossistemas, oferecendo uma gama de funções vitais. Destacam-se suas capacidades na produção de alimentos e na absorção e filtragem da água da chuva, contribuindo para reabastecer os lençóis freáticos e mitigar o risco de inundações e erosão [Ferreira et al. 2023].

Além disso, o solo 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. Contudo, 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. Portanto, 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].

A importância do estoque de carbono orgânico no solo (SOCs) na regulação das mudanças climáticas tem impulsionado a demanda por técnicas que auxiliem na determinação dessa característica [Szatmari et al. 2023, Ye et al. 2021]. Entre essas técnicas, destacam-se os algoritmos de aprendizado de máquina. Nos últimos anos, tem havido um aumento significativo nas atividades de mapeamento de solos, impulsionado pela demanda por informações quantitativas e espaciais [Haddad et al. 2018, Wadoux et al. 2020].

Os algoritmos de Aprendizado de Máquina (AM), por não estarem condicionados a seguir nenhuma suposição estatística, geralmente parecem mais precisos que os modelos estatísticos convencionais, lidando bem com dados massivos e altamente multicolineares. Por este motivo, muitos trabalhos têm focado na aplicação de AM como modelos caixa-preta [Wadoux et al. 2020], sem uma atenção significativa à qualidade e ao tratamento dos dados utilizados no treinamento de tais modelos [Kumar et al. 2023].

Neste sentido, o presente trabalho visa utilizar uma abordagem centrada em dados, comparando diferentes técnicas de AM para estimar a quantidade de estoque de carbono orgânico no solo (SOCs) no território brasileiro. Os modelos desenvolvidos foram avaliados pelas métricas de Média dos Erros Absolutos (MAE) e da Raiz do Erro Quadrático Médio (RMSE). Uma avaliação com dados da Amazônia Central foi realizada e os resultados obtidos mostram que a abordagem desenvolvida consegue ser mais precisa quando comparada com abordagens centradas no modelo [Ferreira et al. 2023].

O artigo está organizado em mais 4 seções. A Seção 2 apresenta uma breve discussão sobre os trabalhos relacionados. A Seção 3 descreve a metodologia. Os resultados são discutidos na Seção 4 e, por fim, a Seção 5 apresenta as considerações finais.

2. Trabalhos Relacionados

Com o crescente volume de dados gerados e com uma maior complexidade dos problemas da área de ciência do solo, o uso de técnicas de Aprendizado de Máquina (AM) cresceu significativamente [Wadoux et al. 2020]. A maioria dos trabalhos de ciência do solo foca na utilização de técnicas de AM como modelos caixa-preta, de certa forma negligenciado tanto os aspectos do pré-processamento de dados, como seleção de atributos, redução de dimensionalidade, validação cruzada, entre outros, quanto as técnicas para otimizar os hiperparâmetros dos modelos.

No contexto brasileiro, diversos trabalhos foram desenvolvidos para avaliar a quantidade de estoque de carbono em solos (SOCs) usando técnicas estatísticas [Ceddia et al. 2015, Ceddia et al. 2016]. Tais trabalhos utilizavam funções de pedotransferência para determinar uma estimativa de densidade do solo e, posteriormente, determinar o SOCs. Essas funções consistem em modelos físico-matemáticos que permitem determinar variáveis complexas a partir de variáveis mais facilmente encontradas, no entanto são técnicas que necessitam de premissas em relação a distribuição dos dados, o que dificulta sua utilização [Tranter et al. 2007, Al-Qinna and Jaber 2013].

Com a capacidade dos métodos de aprendizado de máquina de trabalharem com

dados sem nenhuma suposição estatística *a priori*, alguns trabalhos começaram a ser desenvolvidos comparando essas técnicas com as funções de pedotransferência. Em [Haddad et al. 2017], os autores compararam o desempenho de uma função de pedotransferência com uma rede neural (RN) para estimar a densidade do solo e, posteriormente, o SOCS na região da Amazônia Central com resultados superiores para a técnica de RN. Em [Haddad et al. 2018] os autores utilizaram um comitê de regressores neurais para estimar a densidade do solo.

Em [Gomes et al. 2019] os autores utilizaram diversas técnicas de AM para determinar o SOCs em diferentes regiões do Brasil. Os autores indicaram que os melhores resultados foram obtidos por meio da técnica de *Random Forest* (RF), corroborando com os resultados da revisão de literatura apresentada em [Wadoux et al. 2020] sobre o uso intenso desta técnica na área de ciência do solo. Além disso, os autores identificam que a região que possui o maior valor de SOCs no território brasileiro é a região da Amazônia.

Uma revisão detalhada sobre os modelos de AM usados na determinação de estoque de carbono no solo no Brasil pode ser vista em [Ferreira et al. 2023]. Este estudo explorou a previsão de SOC nas áreas remotas da Amazônia com o uso de 21 covariáveis. O estudo comparou os desempenhos por meio da MAE e do RMSE com os algoritmos de árvore de regressão (RT) e RF. Os resultados mostraram que os modelos de AM tiveram maior precisão na camada de 100cm, onde provavelmente, segue padrões mais estruturais da paisagem representados pelas covariáveis.

A partir da busca por trabalhos relacionados, não foram encontrados trabalhos no domínio de solos que realizassem uma investigação centrada em dados associada a otimização de hiperparâmetros no contexto brasileiro. Assim, este trabalho realiza uma introdução neste assunto, realizando um pré-processamento dos dados associado com técnicas de determinação de hiperparâmetros no contexto da determinação de estoque de carbono no solo (SOCs) nas regiões brasileiras.

3. Metodologia

A Figura 1 ilustra a metodologia adotada: base de dados, representando a coleta e análise de dados; pré-processamento dos dados; seleção de atributos; treinamento do algoritmo de AM supervisionado associado com a técnica de otimização dos hiperparâmetros; e a avaliação de desempenho dos modelos.

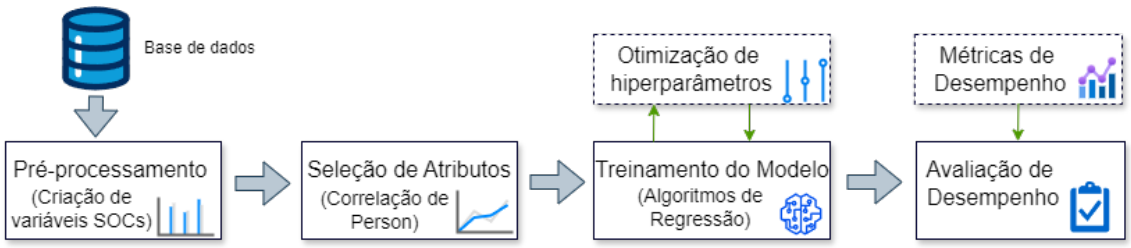

Figura 1. *Pipeline* da metodologia adotada.

O conjunto de dados utilizado nesta pesquisa é baseado no trabalho anterior realizado por [Haddad et al. 2018]. O conjunto de dados completo consiste de 3.404 registros, abrangendo todas as cinco regiões geográficas do Brasil. Esses registros foram obtidos de

69 fontes distintas de publicações de dados de solo. A inclusão de diversas fontes contribui para uma representatividade mais abrangente dos dados, possibilitando uma modelagem mais representativa. Os dados brutos utilizados foram: *(i)* densidade aparente do solo, *(ii)* areia, *(iii)* silte, *(iv)* argila, *(v)* pH (água), *(vi)* soma de cátions básicos como preditor, *(vii)* acidez provocada pela hidrólise do alumínio (Al^{3+}), *(viii)* hidrogênio, *(ix)* material orgânico no solo (MO) e *(x)* a espessura em relação à superfície (T), totalizando 10 atributos [Haddad et al. 2017, Haddad et al. 2018].

A etapa de pré-processamento dos dados envolveu ainda a inclusão da informação de SOCs na base de dados. Uma forma de determinar o SOCs consiste em calcular o produto da densidade aparente do solo (DS) pela concentração do Material Orgânico no Solo (MO) e pela profundidade do horizonte (espessura em relação à superfície) (T) [Ferreira et al. 2023], conforme descrito na Eq. 1. Esta medida fornece a quantidade de estoque de carbono por unidade de área. Por fim, ainda nesta etapa, os dados numéricos foram normalizados para eliminar a discrepância das unidades de medida entre as variáveis.

$$SOCs = DS * MO * T \tag{1}$$

Onde, SOCs em Kg C m^{-2}, MO em g Kg^{-1}, DS em Mg m^{-3} e T em m.

Na etapa de seleção de atributos, o método estatístico da Correlação de Pearson (CP) foi empregado para avaliar as relações lineares entre os atributos contínuos. O valor desse coeficiente varia entre -1 e 1, onde valores próximos de 1 indicam uma correlação positiva forte, valores próximos de -1 indicam uma correlação negativa forte, e valores próximos de 0 indicam ausência de correlação linear. Essa análise permite que atributos que tenham baixa relação com a variável a ser predita sejam eliminados.

No desenvolvimento dos algoritmos de aprendizado de máquina, observou-se que na área de ciência do solo alguns artigos bem citados utilizavam a técnica de RF [Wadoux et al. 2020, Wadoux et al. 2023, Ferreira et al. 2023]. Tal algoritmo consiste em uma extensão da técnica de árvores de regressão, onde são geradas diversas árvores de tamanho menor e o modelo resultante consiste na média dos valores obtidos por todas as árvores [Silatsa et al. 2020]. Partindo deste princípio, optou-se por avaliar outros algoritmos baseados em árvores.

O *Gradient boosting* (GBoost) é uma técnica que cria várias árvores de decisão sequencialmente geradas a partir dos erros de previsão do modelo da árvore anterior [Mahmoudzadeh et al. 2020]. Outro modelo utilizado foi a Árvore de Decisão (AD), que constrói modelos de regressão na forma de uma estrutura de árvore, dividindo o conjunto de dados em subconjuntos cada vez menores [Mousavi et al. 2022].

Para avaliar os modelos desenvolvidos foram utilizadas as métricas: MAE e RMSE. Essas medidas resumem a diferença média nas unidades de valores observados (y_i) e previstos ($\hat{y}_i$) para o SOCs, conforme descrito nas Eq. 2 e 3, respectivamente.

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

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (3)$$

Os modelos de aprendizado de máquina utilizados possuem diversos hiperparâmetros a serem ajustados antes do processo de treinamento em si. Na área de ciência do solo observou-se que a maioria dos trabalhos utiliza os valores padrão para estes hiperparâmetros [Wadoux et al. 2020, Wadoux et al. 2023, Ferreira et al. 2023]. A otimização de hiperparâmetros consiste no processo de encontrar a combinação certa de valores de hiperparâmetros para obter o máximo desempenho do modelo [Japa et al. 2023].

Na busca pela melhor combinação dos hiperparâmetros, foram avaliadas duas técnicas de otimização associadas com validação cruzada (k-fold=5). O *Grid Search* (GS) é uma técnica que 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 [Song et al. 2022]. O *Random Search* (RS) seleciona aleatoriamente combinações de valores de hiperparâmetros a partir de um espaço de busca predefinido [Japa et al. 2023].

4. Resultados

Os experimentos foram realizados em uma máquina Dell Inspiron Intel Core i7-1255U CPU 3.20GHz, com 10 núcleos e 16GB de memória RAM. Os códigos foram implementados em Python versão 3 com as bibliotecas *pandas*[1], *numpy*[2] e *scikit-learn*[3].

A partir do resultado da técnica da Correlação de Pearson, decidimos pela exclusão do atributo "silte" pois foi o menos relevante para a criação de um modelo preditivo. A Tabela 1 apresenta o resultado para SOCs com as técnicas de otimização de hiperparâmetros. Nela, pode-se observar que, para o estoque de carbono orgânico, os dois métodos de otimização de hiperparâmetro associados ao GBoost obtiveram melhores resultados nas duas métricas de desempenho avaliadas em relação aos métodos AD e RF. Além disso, o tempo computacional para o GBoost e para a AD foram equivalentes enquanto para a técnica de RF este tempo foi três vezes maior.

Tabela 1. Resultado para SOCs e otimização de hiperparâmetro com dados de todas as regiões do Brasil.

Variável alvo	Modelo de AM	Otimizacão de Hiperparâmetro	RMSE	MAE
SOCs	RF	RS	0,26	0,17
	RF	GS	0,26	0,17
	GBoost	RS	0,19	0,13
	GBoost	GS	0,19	0,13
	AD	RS	0,37	0,26
	AD	GS	0,37	0,26

[1]https://pandas.pydata.org/
[2]https://numpy.org/
[3]https://scikit-learn.org/stable/

Os modelos gerados foram comparados com os resultados do artigo [Ferreira et al. 2023]. Tal escolha deve-se tanto ao fato da importância ambiental da Amazônia Central, quanto o fato de os autores utilizarem a técnica RF sem técnicas de pré-processamento e otimização dos hiperparâmetros, o que auxilia na comparação. No entanto, os autores fizeram avaliações em diferentes níveis de profundidade do solo, 0-30cm e 0-100cm (SOCS30 e SOCS100), pois a concentração de carbono no solo é maior nas camadas mais superficiais. Observando que os dados de SOCS100 compreendem os dados de SOCS30 e mais 32% de dados, neste trabalho optou-se por avaliar os dados de solos referentes até 100cm de espessura.

A Tabela 2 apresenta a comparação dos resultados obtidos pela técnica de RF desenvolvida aqui com os resultados obtidos em [Ferreira et al. 2023]. Observa-se que os resultados deste trabalho obtiveram melhores valores tanto para o RMSE quanto para o MAE, indicando que um tratamento adequado dos dados com técnicas de pré-processamento e que a otimização dos hiperparâmetros podem auxiliar no desenvolvimento de modelos mais precisos.

Tabela 2. Comparação das métricas RMSE e MAE do modelo RF com o artigo referente à Amazônia Central.

Modelo de AM	Variável alvo	RMSE	MAE
RF	SOCS100 [Ferreira et al. 2023]	2,48	1,90
RF	SOCS (resultado)	0,26	0,17

5. Considerações Finais

O presente trabalho demonstra a eficácia das técnicas de aprendizado de máquina, especialmente RF, GBoost e AD, na estimativa do estoque de carbono orgânico do solo (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 precisão dos modelos preditivos. Em particular, os modelos desenvolvidos superaram as abordagens existentes na literatura para a região da Amazônia Central [Ferreira et al. 2023].

Embora a técnica RF tenha mostrado bons resultados de previsão, 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. A complexidade inerente do RF destaca a importância de explorar novas técnicas e abordagens que possam oferecer uma interpretação mais clara dos resultados. Para trabalhos futuros, propomos testar outros algoritmos de aprendizado de máquina 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 precisão e a aplicabilidade das estimativas de SOCs.

Referências

Al-Qinna, M. and Jaber, S. (2013). Predicting soil bulk density using advanced pedotransfer functions in an arid environment. *Transactions of the ASABE*, 56(3):963–976.

Ceddia, M. B. et al. (2015). Spatial variability of soil carbon stock in the urucu river basin, central amazon-brazil. *Science of the Total Environment*.

Ceddia, M. B. et al. (2016). The use of pedotransfer functions and the estimation of carbon stock in the central amazon region. *Scientia Agricola*.

Ferreira, A. C. S. et al. (2023). Predicting soil carbon stock in remote areas of the central amazon region using machine learning techniques. *Geoderma Regional - Elsevier*.

Gomes, L. C. et al. (2019). Modelling and mapping soil organic carbon stocks in brazil. *Geoderma*, 340:337–350.

Haddad, D. B. et al. (2017). 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*.

Haddad, D. B. et al. (2018). Brazilian soil bulk density prediction based on a committee of neural regressors. *IEEE - Institute of Electrical and Electronic Engineers*.

Japa, L. et al. (2023). A population-based hybrid approach for hyperparameter optimization of neural networks. *IEEE Access*, 11:50752–50768.

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

Mahmoudzadeh, H. et al. (2020). Spatial prediction of soil organic carbon using machine learning techniques in western iran. *Geoderma Regional - Elsevier*.

Mousavi, S. R. et al. (2022). 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*.

Silatsa, F. B. et al. (2020). Assessing countrywide soil organic carbon stock using hybrid machine learning modelling and legacy soil data in cameroon. *Geoderma - Elsevier*.

Song, J. et al. (2022). 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*.

Szatmari, G. et al. (2023). Countrywide mapping and assessment of organic carbon saturation in the topsoil using machine learning-based pedotransfer function with uncertainty propagation. *Catena - Elsevier*.

Tranter, G. et al. (2007). Building and testing conceptual and empirical models for predicting soil bulk density. *Soil Use and Management*, 23(4):437–443.

Wadoux, A. M.-C. et al. (2020). Machine learning for digital soil mapping: applications, challenges and suggested solutions. *Earth-Science Reviews - Elsevier*.

Wadoux, A. M. J.-C. et al. (2023). Shapley values reveal the drivers of soil organic carbon stock prediction. *Soil*.

Ye, Z. et al. (2021). 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*.