UNIVERSIDADE ESTADUAL DE PONTA GROSSA

SETOR DE ENGENHARIAS, CIÊNCIAS AGRÁRIAS E DE TECNOLOGIA

PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO APLICADA

DANILO ALVES DA ROSA

ABORDAGENS DE APRENDIZADO DE MÁQUINA EM DADOS ESPECTRAIS PARA ESTIMATIVA DE ELEMENTOS QUÍMICOS NO SOLO

PONTA GROSSA

2026

DANILO ALVES DA ROSA

ABORDAGENS DE APRENDIZADO DE MÁQUINA EM DADOS ESPECTRAIS PARA ESTIMATIVA DE ELEMENTOS QUÍMICOS NO SOLO

Dissertação apresentada ao Programa de Pós-Graduação em Computação Aplicada, curso de Mestrado em Computação Aplicada da Universidade Estadual de Ponta Grossa, como requisito parcial para obtenção do título de Mestre em Computação Aplicada.

Orientador: Prof.ª Dr.ª Alaine Margarete Guimarães
Coorientador: Prof. Dr. Eduardo Fávero Caires

PONTA GROSSA
2026

Rosa, Danilo Alves da

R788 Abordagens de aprendizado de máquina em dados espectrais para estimativa de elementos químicos no solo / Danilo Alves da Rosa. Ponta Grossa, 2026.

69 f.

Dissertação (Mestrado em Computação Aplicada - Área de Concentração: Computação para Tecnologias em Agricultura), Universidade Estadual de Ponta Grossa.

Orientadora: Profa. Dra. Alaine Margarete Guimarães.
Coorientador: Prof. Dr. Eduardo Fávero Caires.

1. Solo - Análise química. 2. Espectroscopia de infravermelho próximo. 3. Carbono orgânico do solo - Estimativa. 4. Nitrogênio - Solo - Determinação. 5. Agricultura de precisão. I. Guimarães, Alaine Margarete. II. Caires, Eduardo Fávero. III. Universidade Estadual de Ponta Grossa. Computação para Tecnologias em Agricultura. IV.T.

CDD: 631.417

Ficha catalográfica elaborada por Vera Lucia Braga da Silva- CRB9/1026

UNIVERSIDADE ESTADUAL DE PONTA GROSSA
Av. General Carlos Cavalcanti, 4748 - Bairro Uvaranas - CEP 84030-900 - Ponta Grossa - PR - https://uepg.br

TERMO

TERMO DE APROVAÇÃO
Danilo Alves da Rosa

"ABORDAGENS DE APRENDIZADO DE MÁQUINA EM DADOS ESPECTRAIS PARA ESTIMATIVA DE ELEMENTOS QUÍMICOS NO SOLO"

Dissertação aprovada como requisito parcial para obtenção do grau de Mestre no Programa de Pós-Graduação em Computação Aplicada da Universidade Estadual de Ponta Grossa, pela seguinte banca examinadora:

Profa. Dra. Alaine Margarete Guimarães (UEPG/PR - Presidente)

Prof. Dr. José Carlos Ferreira da Rocha (UEPG-PR)

Profa. Dra. Mauren Louise Sguario Coelho de Andrade (UTFPR-PG)

Ponta Grossa, 09 de março de 2026.

Documento assinado eletronicamente por **Alaine Margarete Guimaraes**, **Professor(a)**, em 09/03/2026, às 11:12, conforme Resolução UEPG CA 114/2018 e art. 1º, III, "b", da Lei 11.419/2006.

Documento assinado eletronicamente por **Jose Carlos Ferreira da Rocha**, **Coordenador(a) do Programa de Pós-Graduação em Computação Aplicada - Mestrado**, em 09/03/2026, às 11:12, conforme Resolução UEPG CA 114/2018 e art. 1º, III, "b", da Lei 11.419/2006.

Documento assinado eletronicamente por **MAUREN LOUISE SGUARIO COELHO DE ANDRADE**, **Usuário Externo**, em 09/03/2026, às 15:24, conforme Resolução UEPG CA 114/2018 e art. 1º, III, "b", da Lei 11.419/2006.

A autenticidade do documento pode ser conferida no site https://sei.uepg.br/autenticidade informando o código verificador **3021372** e o código CRC **4C4A9C96**.

AGRADECIMENTOS

Primeiramente, agradeço a Deus, pela força, paciência e sabedoria ao longo desta jornada.

Aos meus familiares, pelo apoio constante e compreensão nos momentos de ausência e dedicação intensa aos estudos. Em especial, aos meus pais, que sempre acreditaram em mim e me incentivaram a lutar pelos meus sonhos.

À minha orientadora, pelo conhecimento compartilhado, paciência, valiosas orientações e incentivo contínuo durante toda a realização deste trabalho. Suas contribuições foram fundamentais para o desenvolvimento desta pesquisa.

Aos professores e colegas do curso de mestrado, pela troca de experiências, debates construtivos e apoio acadêmico.

Agradeço à CAPES – Comissão de Aperfeiçoamento de Pessoal de Nível Superior – pela concessão de bolsas de estudo.

Por fim, agradeço aos amigos que, de alguma forma, contribuíram para minha jornada acadêmica, seja com palavras de incentivo, momentos de descontração ou apoio moral.
A todos que, direta ou indiretamente, fizeram parte desta caminhada, meu sincero, muito obrigado.

RESUMO

A necessidade de métodos mais eficientes e sustentáveis para a análise de componentes químicos do solo impulsiona a integração de tecnologias inovadoras na agricultura. Este trabalho avaliou o potencial de diferentes abordagens de aprendizado de máquina (AM) para estimar os teores de carbono orgânico total (COT) e de nitrogênio total (N) em solos agrícolas, utilizando dados de espectroscopia na região do infravermelho próximo e considerando transformações aplicadas a esses mesmos dados para aprimorar o potencial preditivo dos algoritmos. Foram utilizadas 235 amostras de solo do Laboratório de Fertilidade do Solo da UEPG, com dados espectrais na faixa de 1.100 a 1.648 nm, totalizando 277 atributos de predição. A metodologia incluiu o pré-processamento dos dados, com transformação logarítmica, padronização (Z-score), normalização (Min-Max Scaler) e aplicação do filtro de Savitzky-Golay. Foram empregados sete algoritmos de AM, cujos desempenhos foram avaliados em duas abordagens: utilizando todos os atributos espectrais e após a aplicação da técnica de seleção de atributos SPA (Successive Projections Algorithm). Os resultados indicaram que os modelos Partial Least Squares Regression (PLSR) e Orthogonal Matching Pursuit (OMP), robustos à multicolinearidade, apresentaram desempenho superior, com R^2 de até 0,9124 para N e 0,8156 para COT, ao utilizar a base de dados completa transformada por meio da associação sucessiva dos métodos de transformação logarítmica, de padronização (Z-score) e de filtro de Savitzky-Golay. O uso de seleção de atributos pelo método SPA melhorou significativamente o desempenho da Regressão Linear ($R^2 = 0,8975$) na estimativa de N, porém não proporcionou melhorias quando associado ao algoritmo PLSR. A análise de importância das variáveis no modelo PLSR identificou que as faixas espectrais entre 1.394 nm e 1.412 nm foram as mais relevantes para a predição de N, enquanto, para o COT, as faixas mais importantes situaram-se entre 1.396 nm e 1.450 nm. Conclui-se que: (i) a combinação de espectroscopia NIR com AM, por meio do uso em conjunto dos métodos de transformação logarítmica, padronização (Z-score) e filtro Savitzky-Golay, e associados aos algoritmos PLSR e OMP, consiste em uma metodologia promissora para a estimativa do teor de COT e N no solo, (ii) é possível gerar modelos simples baseados em regressão linear para estimativa de N no solo, quando o método de seleção de atributos SAP é utilizado previamente, (iii) a faixa espectral entre 1.394 nm e 1.450 nm é de especial interesse para a predição de N e COT.

Palavras-chave: Infravermelho Próximo, Filtro Savitzky-Golay, Busca de Correspondência Ortogonal, Carbono Orgânico Total, Nitrogênio no Solo.

ABSTRACT

The need for more efficient, sustainable methods to analyze soil chemical components is driving the integration of innovative technologies in agriculture. This study evaluated the potential of different machine learning (ML) approaches for estimating Total Organic Carbon (TOC) and Total Nitrogen (N) in agricultural soils, using near-infrared spectroscopy data and considering transformations to improve the algorithms' predictive performance. A total of 235 soil samples from the UEPG Soil Fertility Laboratory were used, with spectral data in the 1,100–1,648 nm range, comprising 277 predictive attributes. The methodology included data preprocessing with logarithmic transformation, standardization (Z-score), normalization (Min-Max Scaler), and a Savitzky-Golay filter. Seven ML algorithms were used, and their performance was evaluated in two approaches: using all spectral attributes, and after applying the SPA (Successive Projections Algorithm) attribute selection technique. The results indicated that the models Partial Least Squares Regression (PLSR) and Orthogonal Matching Pursuit (OMP), which are robust to multicollinearity, presented superior performance, with R^2 of up to 0.9124 for N and 0.8156 for TOC, when using the complete database transformed by the successive combination of logarithmic transformation, standardization (Z-score), and Savitzky-Golay filters. The use of feature selection by the SPA method significantly improved the performance of Linear Regression (R^2 = 0.8975) for N estimation, but did not bring improvements when associated with the PLSR algorithm. The analysis of variable importance in the PLSR model identified that the spectral bands between 1,394 nm and 1,412 nm were the most relevant for N prediction, while for TOC, the most important bands were between 1,396 nm and 1,450 nm. It is concluded that: (i) the combination of NIR spectroscopy with ML, through the joint use of logarithmic transformation methods, standardization (Z-score) and Savitzky-Golay filters, and associated with the PLSR and OMP algorithms, consists of a promising methodology for the analysis of TOC and N content in the soil, (ii) it is possible to generate simple models based on linear regression for estimating N in the soil, when the SAP attribute selection method is previously used, (iii) the spectral range between 1,394 nm and 1,450 nm is of special interest for the prediction of N and TOC.

Keywords: Near Infrared, Savitzky-Golay filter, Orthogonal Matching Pursuit, Total Organic Carbon, Nitrogen in soil.

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE SIGLAS

(ABC) Centro de Pesquisa para Assistência e Difusão Técnica Agropecuária da Fundação

(AM) Aprendizado de Máquina

(ANN) *Artificial Neural Network*

(B) Boro

(CLABMU) Complexo de Laboratórios Multiusuários da UEPG

(CNN) *Convolutional Neural Network*

(COT) Carbono Orgânico Total

(DL) *Deep Learning*

(DrSeqANN) *Deep Sequential Artificial Neural Network*

(ELM) *Extreme Learning Machine*

(IA) Inteligência Artificial

(JCR) *Journal Citation Reports*

(LIBS) *Laser-Induced Breakdown Spectroscopy*

(MAE) *Mean Absolute Error*

(MAPE) *Mean Absolute Percentual Error*

(MARS) *Multivariate Adaptive Regression Splines*

(MIR) *Mid-infrared*

(N) Nitrogênio Total

(NIR) *Near Infrared Spectroscopy*

(OLSE) *Ordinary Least Squares Estimation*

(PLSR) *Partial Least Squares Regression*

(R^2) Coeficiente de Determinação

(RF) *Random Forest*

(RMSE) *Root Mean-Square Error*

(S) Enxofre

(SG) Savitzky-Golay

(SPA) *Successive Projections Algorithm*

(SVM) *Support Vector Machine*

(UEPG) Universidade Estadual de Ponta Grossa

(WD-XRF) *Wavelength Dispersive X-ray Fluorescence*

(WOA) *Whale Optimization Algorithm*

(XANES) *X-ray Absorption Near-Edge Structure*

SUMÁRIO

1 INTRODUÇÃO

A agricultura moderna enfrenta desafios cada vez maiores para aumentar a produtividade, minimizar os impactos ambientais e assegurar a sustentabilidade dos sistemas de produção agrícola. Neste contexto, o acompanhamento minucioso dos componentes químicos do solo, como o carbono orgânico total (COT) e o nitrogênio total (N), é crucial para orientar decisões agronômicas, uma vez que esses elementos desempenham funções indispensáveis à fertilidade do solo e ao crescimento das plantas.

A espectroscopia na região do infravermelho próximo (NIR, do inglês *Near Infrared*) tem sido uma alternativa para estimar teores de nutrientes no solo, oferecendo técnicas rápidas, não destrutivas e capazes de gerar dados multivariados em tempo real (Ahmadi et al., 2021). Contudo, os extensos dados espectrais do solo, caracterizados por alta dimensionalidade de banda, normalmente contêm informações inválidas, redundantes e sobrepostas. Essa complexidade apresenta desafios, gerando instabilidade e dificultando os esforços para aprimorar a precisão dos modelos de predição (He et al., 2023).

A interpretação de espectros complexos exige abordagens computacionais avançadas para extrair padrões que correlacionem as bandas espectrais com as características do solo em análise. Nesse sentido, o Aprendizado de Máquina (AM) se destaca por permitir a modelagem de relações entre atributos, a seleção de características relevantes e a predição de parâmetros físico-químicos com alta acurácia. O sucesso dos processos de AM depende da qualidade e das características dos dados a serem processados. Dessa forma, a etapa de pré-processamento, em que os dados são tratados e, quando necessário, transformados, é determinante para a qualidade dos modelos finais gerados.

Diante desse cenário, este trabalho investiga o uso de dados espectrais NIR associados a técnicas de pré-processamento e algoritmos de aprendizado de máquina para a estimativa do teor de carbono orgânico total (COT) e nitrogênio (N) em solos agrícolas. Para isso, inicialmente são apresentados os objetivos da pesquisa na seção 2. Em seguida, é realizada uma revisão da literatura abordando conceitos relacionados à espectroscopia, técnicas de pré-processamento de dados espectrais e métodos de aprendizado de máquina aplicados à análise de solos na seção 3. Posteriormente, são descritos os materiais e métodos utilizados no desenvolvimento do estudo, incluindo a base de dados, os procedimentos de pré-processamento e os algoritmos empregados na construção dos modelos preditivos na seção 4. Na sequência, são apresentados e discutidos os resultados obtidos, destacando o desempenho das diferentes

abordagens avaliadas na seção 5. Por fim, são apresentadas as conclusões do trabalho, bem como as principais contribuições e perspectivas para pesquisas futuras na seção 6.

2 OBJETIVOS

2.1 Objetivo Geral

Estimar o teor de COT e de N em solos agrícolas, a partir de dados do espectro NIR, associando métodos de transformação de dados a distintas abordagens algorítmicas de Aprendizado de Máquina.

2.2 Objetivos Específicos

1. Realizar o pré-processamento da base de dados, considerando os métodos mais adequados para o tratamento e a transformação de dados do espectro NIR.
2. Gerar e validar modelos de estimativa de COT e N, utilizando diferentes abordagens algorítmicas de AM.
3. Analisar o desempenho dos métodos empregados por meio de métricas estatísticas de avaliação.

3 REVISÃO DE LITERATURA

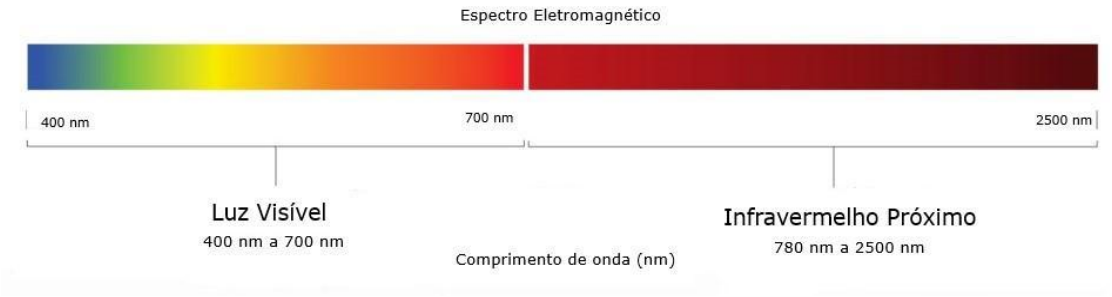
3.1 Espectroscopia

A espectroscopia é uma técnica analítica que estuda a interação da luz com a matéria. Ela permite a identificação e quantificação de compostos químicos com base nas características de absorção, emissão ou espalhamento de radiação por esses compostos em diferentes comprimentos de onda.

No contexto da análise de solo, a espectroscopia é usada para identificar e quantificar nutrientes e compostos orgânicos e inorgânicos presentes nas amostras de solo, por meio da análise de sua resposta a diferentes tipos de radiação.

A Figura 1 apresenta uma representação do espectro eletromagnético, demonstrando os diferentes comprimentos de onda.

Figura 1. Espectro eletromagnético



Fonte: O autor.

A espectroscopia envolve a passagem de radiação (geralmente luz, em várias faixas do espectro eletromagnético, como ultravioleta, visível, infravermelho, infravermelho próximo, etc.) através de uma amostra ou a medição da radiação que a amostra emite ou espalha. Dependendo da interação da radiação com a amostra, é possível obter informações sobre a composição química, a estrutura molecular, as propriedades físicas e até a concentração de substâncias presentes (Mirani et al., 2021).

Ao observar o espectro, notam-se diferentes intervalos (faixas) de comprimento de onda que podem ser usados para diferentes propósitos, como micro-ondas, raio X, ultravioleta, infravermelho e, dentro deste, o infravermelho próximo.

3.1.1 Espectroscopia na região do infravermelho próximo

A espectroscopia NIR é utilizada na agricultura com várias finalidades, entre elas a detecção de nutrientes, como nitrogênio, enxofre e matéria orgânica no solo (Santos et al., 2020).

Vários estudos têm utilizado a espectroscopia NIR para estimar o teor de N no solo, com a vantagem de ser uma técnica não destrutiva e de rápida execução (Pasquini, 2018). No entanto, a presença de compostos orgânicos e minerais no solo pode interferir nas medições, dificultando a obtenção de estimativas precisas sem calibração rigorosa e constante.

A espectroscopia NIR tem se destacado por sua capacidade de quantificar o teor de COT no solo, por meio de bandas específicas (Tamburini et al., 2017). A estimativa do COT é um parâmetro importante, pois está diretamente relacionada à fertilidade do solo e à capacidade de retenção de água. Porém, assim como para o N, ainda não existe uma metodologia clara de tratamento de dados NIR, tampouco um modelo de estimativa eficiente e efetivamente disponível para uso na estimativa de COT.

3.2 Pré-processamento de Dados Espectrais

O pré-processamento de dados tem como propósito principal minimizar a variabilidade não explicada pelos modelos, de modo a evidenciar, nos espectros, a característica de interesse, geralmente relacionada a um fenômeno ou componente específico. Quando bem escolhida, a técnica de pré-processamento pode atingir esse propósito; entretanto, há sempre o risco de utilizar um método inadequado ou excessivamente rigoroso, que acabe eliminando informações importantes. Definir qual é a melhor técnica de pré-processamento a ser adotada para determinada base de dados é uma tarefa que só pode ser confirmada por meio da validação do modelo.

O objetivo da etapa de pré-processamento na análise de dados espectrais pode ser: (i) melhorar uma análise exploratória subsequente, (ii) melhorar um modelo de calibração subsequente forçando os dados a obedecer à lei de Lambert-Beer, que é empírica para dados espectrais NIR e sugere uma relação linear entre a absorbância dos espectros e a concentração

do elemento químico analisado; ou ainda (iii) melhorar um modelo de estimativa subsequente (Rinnan, 2009).

Os métodos de pré-processamento mais amplamente utilizados em espectroscopia NIR (tanto no modo de refletância quanto no de transmitância) podem ser divididos em duas categorias: métodos de correção de dispersão e derivadas espectrais. O grupo de métodos de correção de dispersão inclui a correção de espalhamento multiplicativa, a remoção de tendência, a variável normal padrão (SNV) e a normalização. O grupo de derivação espectral é representado por técnicas como as derivadas de Norris-Williams (NW) e os filtros de derivadas polinomiais de Savitzky-Golay (SG) (Rinnan, 2009).

Para esse trabalho, serão considerados: a normalização, como método de correção de dispersão, e os filtros SG, como método de derivação espectral.

3.2.1 Normalização de dados

O uso de métodos de normalização pode proporcionar benefícios consideráveis na detecção de anomalias na distribuição de dados e no uso do modelo em estudo, de forma mais eficaz e eficiente (Akilli e Atil, 2020).

A normalização Mín-Máx fornece uma transformação linear dos dados originais, mapeando os valores dos atributos para um intervalo de 0 a 1 ou de -1 a 1. A vantagem da normalização Mín-Máx é preservar a relação entre os dados originais; ou seja, mantém a distribuição dos dados, preservando as relações de proporcionalidade entre os valores originais. No entanto, esse método é altamente afetado por valores extremos ou outliers, pois os resultados podem ser dominados por valores específicos elevados (Jain et al., 2005 apud Akilli e Atil, 2020).

O método Min-Max reescalona cada valor do atributo de acordo com a fórmula apresentada na Figura 2.

Figura 2. Fórmula transformação linear MinMaxScaler

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}}$$

Fonte: Amorim et al. (2023)

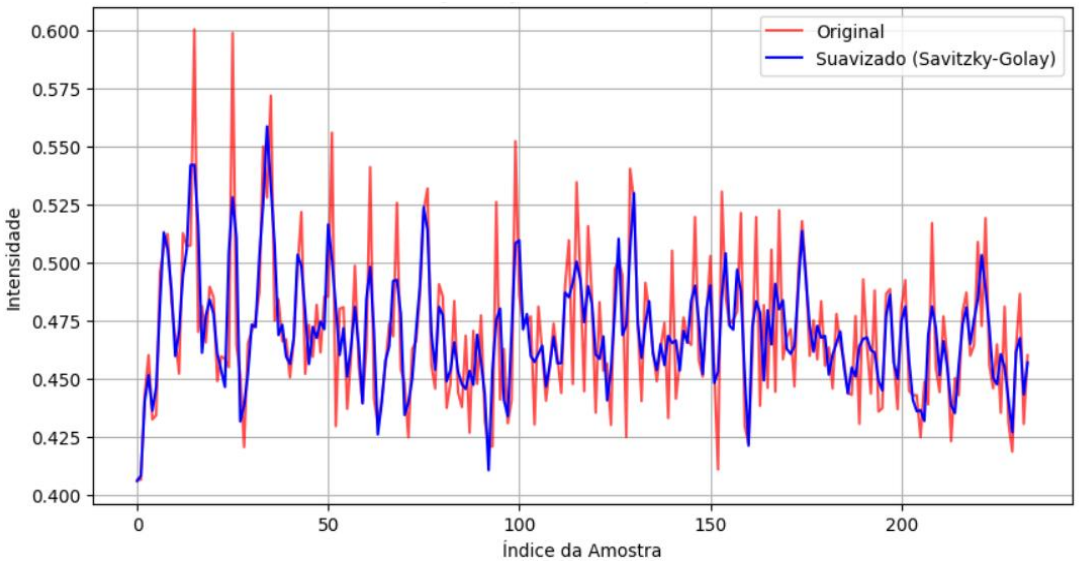
Na equação apresentada na Figura 2, x representa o valor original, x_{min} e x_{max} são, respectivamente, o menor e o maior valor daquela variável, e x' é o valor transformado. Dessa forma, o menor valor do atributo passa a ser representado por 0, o maior por 1, e os demais são ajustados proporcionalmente nesse intervalo.

3.2.2 Filtro Savitzky-Golay

O filtro Savitzky-Golay (SG) é um método de derivação espectral que utiliza a suavização dos espectros antes do cálculo da derivada, a fim de reduzir o efeito prejudicial na relação sinal-ruído que as derivadas convencionais de diferenças finitas apresentam.

Esse filtro é usado para suavizar os dados, sem distorcer muito os picos nem a forma geral do sinal. Ele é especialmente útil em dados espectrais, pois é um filtro de média móvel polinomial. Em vez de apenas tirar a média simples dos pontos vizinhos, ele ajusta um polinômio a uma janela deslizante dos dados e usa-o para calcular um valor suavizado no ponto central (Schmid et al.,2022). Na Figura 3 tem-se um exemplo do efeito da aplicação do filtro SG, em que o ruído do espectro (linha azul) foi reduzido, mantendo a forma e a altura dos picos originais (linha vermelha).

Figura 3. Filtro Savitzky-Golay



Fonte: O autor.

3.2.3 Transformação logarítmica

A transformação logarítmica é uma técnica amplamente utilizada no pré-processamento de dados para AM, especialmente em estudos sobre nutrientes do solo. Isso se deve ao fato de que variáveis como N e COT costumam apresentar distribuições assimétricas, com muitos valores baixos e poucos muito altos. Além disso, diferentes nutrientes podem estar em escalas distintas, o que dificulta a comparação direta, e suas variações geralmente são multiplicativas.

A aplicação do logaritmo aos dados ajuda a suavizar essas diferenças, reduzindo a influência de valores extremos e aproximando a distribuição de forma mais simétrica. Esse processo melhora a relação entre os atributos de predição e o atributo meta, aproximando-a de uma relação linear, o que pode beneficiar modelos como regressões, redes neurais e SVMs. A forma mais simples de aplicar a transformação logarítmica é apresentada na Figura 4, na qual o logaritmo é aplicado diretamente à variável. Como não é possível calcular o logaritmo de valores menores ou iguais a zero, adiciona-se uma constante à transformação, como apresentado na Figura 5.

Figura 4. Forma mais simples da transformação

$$X_{\log} = \log(X)$$

Figura 5. Forma da transformação logarítmica acrescida de uma constante

$$X_{\log} = \log(X + c)$$

Fonte: O autor.

Entre as vantagens dessa abordagem estão a redução da assimetria, a diminuição do impacto de outliers e a melhoria da comparabilidade entre diferentes escalas. No entanto, é preciso ter cuidado na interpretação dos resultados, pois, após a transformação, os valores passam a estar em escala logarítmica. Também é necessário tratar valores negativos ou ajustar adequadamente a constante somada para evitar distorções (Raymaekers et al., 2024).

3.2.4 Padronização

A padronização é um método de redimensionar atributos para que fiquem entre um valor mínimo e um valor máximo, geralmente entre zero e um, ou de modo que o valor absoluto máximo de cada atributo seja dimensionado ao tamanho da unidade.

O Z-*score* (também chamado *score* Z ou escala de desvio-padrão) é um método de padronização de dados que se baseia na média e no desvio-padrão de um atributo. Um Z-escore de zero significa que o valor é exatamente igual à média, enquanto valores positivos ou negativos indicam que o valor está acima ou abaixo da média, respectivamente (YE, 2004).

O Z-escore pode ser útil para comparar dados de diferentes escalas, bem como para identificar valores muito distantes de zero (por exemplo, com Z-escores de +3 ou -3), que podem ser considerados anomalias (outliers) em um conjunto de dados normalmente distribuído.

A detecção e remoção de outliers são utilizadas no pré-processamento de dados, uma vez que valores extremos podem influenciar negativamente análises estatísticas e comprometer a performance de modelos de AM.

O Z-*score* é de fácil implementação, conforme a equação apresentada por Ye (2004) (Figura 6).

Figura 6. Equação de cálculo do Z-score

$$Z = \frac{x - \mu}{\sigma}$$

Fonte: Ye (2004)

Na equação, *x* representa o valor observado, μ é a média da variável e σ é o desvio padrão. Logo, o Z-*score* quantifica a distância de um valor em relação à média em unidades de desvio padrão. Pode-se dizer que valores com Z > 3 são considerados outliers potenciais, por se encontrarem significativamente afastados do comportamento esperado dos dados. Essa técnica pode ser aplicada antes da normalização de dados, pois valores extremos podem distorcer a escala de normalização e comprimir os demais valores próximos de zero (Lartey et al., 2024).

3.2.5 Seleção de Atributos

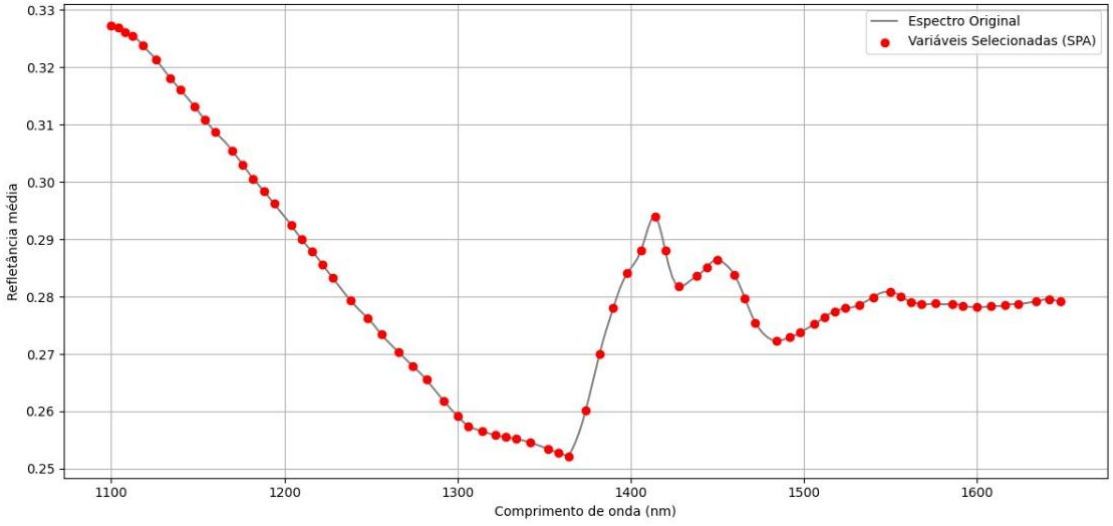
A seleção de atributos pode ser usada para diferentes finalidades; entre as principais, está a redução de dimensionalidade dos dados, uma boa alternativa quando se trata de dados espectrais, visto que são de alta dimensionalidade quando abrangem um amplo espectro.

A seleção de atributos pode ocorrer de forma embebida no algoritmo de AM ou como alternativa a ser executada previamente ao algoritmo de AM. Entre as várias abordagens de seleção de atributos, tem-se o Algoritmo de Projeções Sucessivas (SPA, do *inglês Successive Projections Algorithm*), que pode ser usado para selecionar variáveis espectrais relevantes (como bandas espectrais) em modelos de predição que envolvem dados espectrais e características do solo.

O SPA é um método supervisionado de seleção de variáveis que tem como objetivo reduzir a alta correlação entre elas, selecionando aquelas que fornecem informações complementares ao modelo. Ele seleciona bandas que são importantes para prever o atributo meta, mas não são redundantes entre si.

Um exemplo do uso de SPA é apresentado na Figura 7, na qual o espectro médio (linha cinza) representa os dados originais e os pontos vermelhos são os comprimentos de onda selecionados pelo SPA. Esses pontos estão espalhados de forma a evitar redundância, capturando informações relevantes em diferentes regiões do espectro.

Figura 7. Exemplo de seleção de bandas espectrais utilizando o algoritmo SPA



Fonte: O autor.

3.3 Aprendizado de Máquina Aplicado a Dados Espectrais

O Aprendizado de Máquina é uma subárea da Inteligência Artificial (IA) que permite que os sistemas aprendam a partir de dados e melhorem seus desempenhos ao longo do tempo, sem necessidade de serem programados explicitamente para cada tarefa. Em outras palavras, o AM capacita as máquinas a identificar padrões e a tomar decisões com base em dados, sem intervenção humana direta (Javaid et al., 2023).

Esse processo funciona com base em algoritmos que processam dados e os utilizam para construir modelos. Os modelos são usados para identificar padrões e fazer previsões.

As abordagens de AM têm sido amplamente aplicadas na interpretação de dados espectrais para a predição de propriedades químicas do solo. De acordo com Munawar et al. (2020), técnicas de aprendizado supervisionado permitem modelar relações complexas entre os espectros obtidos por NIR e os teores de nutrientes no solo, superando as limitações dos métodos estatísticos convencionais.

Os algoritmos de AM possuem características próprias que os tornam adequados a diferentes tipos de problemas.

Algoritmos de regressão linear múltipla *(Linear Regression)* são escolhas frequentes na análise de dados espectrais, mas enfrentam o desafio do sobreajuste, especialmente com muitos variáveis e a dificuldade de lidar com dados multicolineares. Como alternativa, a regressão *Ridge*, por meio de técnicas de regularização, pode ser utilizada para aprimorar a precisão do modelo, tratar a multicolinearidade nos dados e evitar o sobreajuste. Esse método impõe penalidades à magnitude dos coeficientes de regressão (Rajan, 2022).

A regressão de mínimos quadrados parciais (PLSR, do inglês *Partial Least Squares Regression*) é um método estatístico que busca modelar a relação entre variáveis preditoras e uma variável resposta, particularmente útil quando há muitas variáveis altamente correlacionadas, como ocorre na espectroscopia. Ao reduzir a dimensionalidade dos dados e extrair componentes latentes, o PLSR melhora a capacidade de predição e de interpretação dos modelos, sendo amplamente aplicado na análise de solos e na composição química de materiais (Pudelko et al., 2020).

A máquina de vetores de suporte (SVM, do inglês *Support Vector Machine*) é uma abordagem supervisionada que pode ser utilizada tanto para a classificação quanto para a regressão. Sua principal característica é encontrar um hiperplano ótimo que separa os dados em diferentes classes ou que melhor ajusta a relação entre variáveis, no caso de regressão. Com o

uso de funções de *kernel*, a SVM é capaz de modelar relações complexas e não lineares, tornando-se uma ferramenta poderosa para análise de padrões em dados agroclimáticos e espectrais (Zhu et al., 2022). Quando aplicada à regressão, a SVM também é denominada SVR (do inglês *Support Vector Regression*).

O algoritmo de Florestas Aleatórias (RF, do inglês *Random Forest*), por sua vez, é um modelo de AM pertencente à categoria de algoritmos *ensemble (agrupados)*. Baseado em múltiplas árvores de decisão construídas a partir de amostras aleatórias dos dados, combina várias predições, que, isoladas, podem ser consideradas mais fracas, para criar um modelo mais forte e preciso. A previsão final da regressão é determinada pela média das previsões. Sua capacidade de lidar com grandes volumes de dados e ruído faz com que o RF seja amplamente empregado na predição de produtividade agrícola e na classificação de solos (Torsoni et al., 2023).

As Redes Neurais Artificiais (ANNs, do inglês *Artificial Neural Networks*) são modelos inspirados no funcionamento do cérebro humano, compostos por camadas de neurônios artificiais interconectados. Esses neurônios processam informações por meio de funções de ativação, permitindo que a rede aprenda padrões complexos e não lineares a partir dos dados (Luo et al., 2022). As ANNs e sua subárea, denominada Aprendizado Profundo (DL, do inglês Deep Learning), são especialmente úteis para modelar fenômenos agroclimáticos e espectrais devido à sua capacidade de generalização e de aprendizado profundo (Jin et al., 2023).

Nos últimos anos, a tecnologia de representação esparsa fez contribuições notáveis em processamento de sinais, de imagens e de outros diferentes tipos de dados provindos de sensores. Algoritmos gulosos são uma classe importante no campo da representação esparsa, sendo o algoritmo de busca de correspondência ortogonal (OMP, do inglês *Orthogonal Matching Pursuit*) um exemplo típico. Considerando as características dos dados espectrais NIR, o OMP pode ser uma alternativa interessante para a análise desses dados (Dong e Wu, 2021).

Cada uma das abordagens algorítmicas de AM citadas tem seus prós e contras, e a seleção depende da complexidade do problema, da quantidade de dados à disposição e da necessidade de interpretação dos resultados. Embora o PLSR seja mais utilizado em situações em que a interpretabilidade é crucial, as abordagens de SVM e RF se destacam pela exatidão nas previsões complexas. Por outro lado, as ANNs são indicadas para a modelagem de dados de alta dimensionalidade e de alta similaridade, em que os padrões não lineares desempenham papel crucial.

3.4 Nitrogênio e Carbono Orgânico Total no Solo

N e COT são fundamentais para o desenvolvimento saudável das plantas e, por isso, sua quantificação precisa no solo é crucial.

O nitrogênio é um nutriente essencial para as plantas e desempenha um papel fundamental na fertilidade do solo. É um componente vital de aminoácidos, que são os blocos de construção das proteínas. Também é um componente essencial de ácidos nucleicos e de clorofila, a molécula responsável pela fotossíntese. A disponibilidade de nitrogênio no solo afeta diretamente o crescimento, a qualidade e a produtividade das plantas (Marschner, 2012).

O carbono orgânico total é uma medida importante que quantifica o total de carbono orgânico presente em uma amostra de solo ou de água, sendo um indicador fundamental da qualidade do solo, relacionado à fertilidade e à capacidade de armazenamento de carbono do solo. Monitorar o COT é crucial para práticas de manejo sustentável e para a implementação da agricultura de precisão, pois afeta diretamente a produtividade e a saúde do solo (Tamburini et al., 2017).

3.5 Trabalhos Correlatos

A Tabela 1 apresenta uma compilação de artigos científicos que utilizaram técnicas de espectroscopia para a análise de nutrientes no solo. Seu objetivo é reunir, de forma organizada, as informações-chave de cada estudo, permitindo uma visão comparativa e abrangente do tema. Entre os dados apresentados estão o título do artigo, o ano de publicação, os nutrientes analisados (como carbono, nitrogênio, enxofre e boro), a classificação Qualis, a presença no *Journal Citation Reports* (JCR), o número de amostras utilizadas, os algoritmos ou métodos estatísticos aplicados, a região geográfica de realização do estudo, as variáveis analisadas além da espectroscopia e os principais resultados obtidos.

Essa estrutura permite identificar tendências e padrões de pesquisa, avaliar quais técnicas espectrais e métodos de modelagem têm apresentado melhor desempenho, além de destacar lacunas que ainda podem ser exploradas em investigações futuras, contribuindo, assim, para o direcionamento metodológico desse trabalho.

Tabela 1. Trabalhos Correlatos

Título Artigos	Ano	Nutriente	Qualis	JCR	Quantidade de Amostras	Algoritmos Utilizados	Região	Variáveis Utilizadas	Resultados obtidos
Soil organic matter characteristics in drained and rewetted peatlands of northern Germany: Chemical and spectroscopic analyses (He, S. et al., 2019).	2019	Nitrogênio e Enxofre	A1	SIM	36	Não informa	Alemanha	Matéria Orgânica do Solo, Carbono Orgânico, Nitrogênio, Enxofre. Umidade Gravimétrica, pH do Solo, Grau de Humificação	Espectroscopia de Absorção de Raios X de Energia (XANES): Especificação de Carbono e Nitrogênio: A análise XANES forneceu informações sobre as diferentes formas químicas de carbono e nitrogênio presentes na MOS. A técnica permitiu a identificação de grupos funcionais específicos, revelando como a drenagem e a reidratação afetaram a humificação e a composição química da MOS
Boron speciation and extractability in temperate and tropical soils: A multi-surface modeling approach (Van et al., 2020).	2020	Boro	A2	SIM	10	CD-MUSIC (*Charge Distribution - Multi-Site Ion Complexation*)	Holanda e Burundi	Tipo de Solo, pH do Solo, Concentrações De Boro, Superfícies Reativas, Métodos de Extração	A espectroscopia de plasma acoplado indutivamente (ICP-OES) e a espectrometria de massa com plasma acoplado indutivamente de alta resolução (HR-ICP-MS). Essas técnicas permitiram a quantificação precisa do boro nas extrações realizadas com soluções como $CaCl_2$ e KH_2PO_4, bem como nas extrações com HNO_3 e manitol

Título Artigos	Ano	Nutriente	Qualis	JCR	Quantidade de Amostras	Algoritmos Utilizados	Região	Variáveis utilizadas	Resultados obtidos
						Quadrados Parciais (PLS)			combinando os dados espectrais com métodos de pré-processamento e algoritmos de regressão para melhorar a precisão das previsões
Sulfur Detection in Soil by Laser Induced Breakdown Spectroscopy Assisted by Multivariate Analysis (Gazeli et al., 2021).	2021	Enxofre	A3	SIM	11	Análise de Componentes Principais (PCA), Análise Multivariada, Análise Univariada	Não informa	Comprimento de onda, Concentração de enxofre, Tempo de atraso	Foram identificadas linhas espectrais do enxofre na região VUV (175–200 nm) que eram suficientemente intensas e relativamente livres de interferências espectrais de outros elementos presentes na matriz do solo. Essas linhas foram selecionadas para medições qualitativas e quantitativas, As linhas espectrais do enxofre na região UV-Vis-NIR (200–1200 nm) foram consideradas inadequadas para análise quantitativa, pois eram muito fracas ou sobrepostas a linhas espectrais de outros elementos, Esses resultados demonstram a viabilidade do uso de LIBS, especialmente na região VUV, para a análise de enxofre em matrizes complexas como o solo orgânico.
Determination of sulfur in Soils and Stream Sediments by Wavelength Dispersive X-ray Fluorescence Spectrometry (Zhao et al., 2020).	2020	Enxofre	A2	SIM	47	Não informa	China	Espécies de Enxofre, Parâmetros Espectrais, Concentrações:	Foi utilizada a Espectroscopia de Fluorescência de Raios X por Dispersão de Comprimento de Onda (WD-XRF) que permitiu a detecção de enxofre em concentrações tão baixas quanto 5,43 mg/kg, demonstrando a sensibilidade da técnica, portanto, a espectroscopia, foi crucial para a análise e determinação precisa do enxofre nas amostras estudadas
Predicting carbon and nitrogen by visible near-infrared (Vis-NIR)	2020	Nitrogênio	A3	SIM	701	Mínimos Quadrados	Brasil	*Multiplicative Scatter*	As previsões de concentrações de carbono (C) e nitrogênio (N) foram mais precisas na região do

Título Artigos	Ano	Nutriente	Qualis	JCR	Quantidade de Amostras	Algoritmos Utilizados	Região	Variáveis utilizadas	Resultados obtidos
and mid-infrared (MIR) spectroscopy in soils of Northeast Brazil (Santos et al.,2020).						Parciais (PLSR), Máquina de Vetores de Suporte (SVM).		*Correction (MSC), StandardNormal Variate* (SNV), Bandas Espectrais	infravermelho médio (MIR) em comparação com a região do visível e infravermelho próximo (Vis- NIR). Não houve melhoria significativa ao combinar as duas regiões espectrais.
Effects of Moisture and Particle Size on Quantitative Determination of Total Organic Carbon (TOC) in Soils Using Near-Infrared Spectroscopy (Tamburini et al., 2017).	2017	TOC	A2	SIM	46	SNV (*Standard Normal* Variate), 2ª Derivada	Itália	Conteúdo de Carbono Orgânico Total (TOC), Umidade do Solo, Textura do Solo, Espectros de Reflectância NIR	Os resultados do estudo indicaram que a espectroscopia no infravermelho próximo (NIR) é uma técnica promissora para prever o carbono orgânico total (TOC) em amostras de solo, independentemente das variações nas características físicas do solo, como tamanho de partícula e conteúdo de umidade.
Predicting the decomposability of arctic tundra soil organic matter with mid infrared spectroscopy (Matamala et al., 2018)	2018	TOC e Nitrogênio	A1	SIM	227	Regressão por Mínimos Quadrados Parciais (PLSR), Análise de Componentes Principais (PCA)	Alasca	Carbono Orgânico Total (TOC), Nitrogênio Total (TN), Carbono Mineralizado, Ambientais.	Os resultados do estudo utilizando espectroscopia de infravermelho médio (MIR) mostraram que essa técnica é eficaz para prever o Carbono Orgânico Total (TOC) e o Nitrogênio Total (TN) em solos de tundra, além de fornecer boas predições para a mineralização de carbono.

Título Artigos	Ano	Nutriente	Qualis	JCR	Quantidade de Amostras	Algoritmos Utilizados	Região	Variáveis utilizadas	Resultados obtidos
Calibration models database of near infrared spectroscopy to predict agricultural soil fertility properties (Munawar et al., 2020)	2020	Nitrogênio	A1	SIM	40	Regressão por Componentes Principais (PCR), Regressão por Mínimos Quadrados Parciais (PLSR)	Indonésia.	Nitrogênio (N) Fósforo (P) Potássio (K) pH do solo Magnésio (Mg) Cálcio (Ca)	Os resultados mostram que os modelos de PLSR geralmente apresentaram um desempenho superior em comparação com os modelos de PCR, especialmente em relação ao fósforo e ao nitrogênio, onde os valores de R^2 e RPD foram mais altos para PLSR
Soil Properties Prediction for Precision Agriculture Using Visible and Near-Infrared Spectroscopy: A Systematic Review and Meta-Analysis (Ahmadi et al., 2021)	2021	TOC, Nitrogênio,	A2	SIM	115	Regressão por Mínimos Quadrados Parciais (PLSR)	China, Estados Unidos, Itália, Alemanha, Brasil.	Refletância do solo: Medida em diferentes comprimentos de onda na região do visível e infravermelho próximo (400 a 2500 nm). Propriedades do solo.	Os valores médios do R^2 para diferentes propriedades do solo variaram de 0,68 a 0,87. Isso sugere que a espectroscopia V-NIR pode prever com precisão propriedades como carbono e nitrogênio, A revisão concluiu que a espectroscopia V-NIR é uma alternativa confiável, de baixo custo e rápida para os métodos padrão de previsão de propriedades do solo, com uma precisão aceitável após mais de 30 anos de pesquisa.

Título Artigos	Ano	Nutriente	Qualis	JCR	Quantidade de Amostras	Algoritmos Utilizados	Região	Variáveis utilizadas	Resultados obtidos
Predictive performance of mobile vis-near infrared spectroscopy for key soil properties at different geographical scales by using spiking and data mining techniques (Nawar et al., 2017)	2017	TOC, Nitrogênio,	A1	SIM	1355	PLS (*Partial Least Squares*), MARS (*Multivariate Adaptive Regression Splines*), SVM (*Support Vector Machine*)	Reino Unido	Matéria Orgânica do Solo (SOM), Textura do Solo, Nutrientes do Solo, Umidade do Solo	Os resultados do estudo indicaram que o método MARS (Multivariate Adaptive Regression Splines) superou os métodos SVM (Support Vector Machine) e PLSR (Partial Least Squares Regression) na modelagem das propriedades do solo durante a validação cruzada. Para o Carbono Total (TC), o MARS também apresentou resultados excelentes, com R^2=0.98 e RMSECV variando entre 0.06% e 0.10% para ambos os campos
Multi-spectral evaluation of total nitrogen, phosphorus and potassium content in soil using Vis-NIR spectroscopy based on a modified support vector machine with whale optimization algorithm (Liu, M. et al., 2025)	2025	Nitrogênio Total (TN), Fósforo Total (TP), Potássio Total (TK)	A1	SIM	99	RBF-poly-SVM (Máquina de Vetor de Suporte com kernel híbrido), *Whale Optimization Algorithm* (WOA), IRIV, VISSA-IRIV (seleção de variáveis)	China	Espectroscopia Visível–Infravermelho Próximo (Vis-NIR) (350–2500 nm)	Para o conteúdo de potássio total (TK), o modelo alcançou um coeficiente de determinação (R^2) de 0.950 no conjunto de calibração e 0.927 na validação independente, com RMSEs de 0.232 e 0.194, respectivamente. A seleção de variáveis espectrais relevantes também contribuiu para a alta precisão, como evidenciado pelos métodos de seleção de variáveis (SPA, VISSA-IRIV, CARS-IRIV).

Título Artigos	Ano	Nutriente	Qualis	JCR	Quantidade de Amostras	Algoritmos Utilizados	Região	Variáveis utilizadas	Resultados obtidos
Prediction of soil organic carbon stock along layers and profiles using Vis-NIR laboratory spectroscopy (Dharumarajan et al., 2025)	2025	TOC	A1	SIM	361	(PLSR)	Índia	As variáveis principais utilizadas no estudo foram relacionadas às propriedades do solo, incluindo o teor de carbono orgânico do solo (SOC) e a densidade do solo (BD)	Os resultados do estudo indicam que a abordagem indireta, que combina a previsão de SOC e densidade do solo (BD) separadamente a partir de espectros Vis-NIR, apresentou maior confiabilidade na estimativa do estoque de carbono orgânico do solo (SOC stock) em comparação com a abordagem direta.
Predicting the Carbon Property of Soil Using DrSeqANN and VIS/NIR Spectroscopy (Jakkan et al., 2025)	2025	Carbono	B1		200	(DrSeq ANN) *Random Forest* (RF) (ELRR)	Índia	As principais variáveis utilizadas na análise foram as características espectrais do solo obtidas na faixa	Especificamente, o modelo DrSeqANN apresentado revelou o melhor desempenho, com valores de RMSE de aproximadamente 0,08, R^2 de 0,82 e RPIQ de 4,32, utilizando os dados pré-processados com Log10x.

Título Artigos	Ano	Nutriente	Qualis	JCR	Quantidade de Amostras	Algoritmos Utilizados	Região	Variáveis utilizadas	Resultados obtidos
								de comprimento de onda de 350 a 2500 nm	
Integrating Fusion Strategies and Calibration Transfer Models to Detect Total Nitrogen of Soil Using Vis-NIR Spectroscopy (Mirani, F. et al., 2025)	2025	Nitrogênio Total	A1	SIM	168	(SVR) (PLSR) (Ridge)	China	As variáveis utilizadas no estudo incluem principalmente o conteúdo de nitrogênio total (TN) no solo, que é o alvo principal de previsão através de espectroscopia Vis-NIR	Para o modelo de previsão de TN aplicado ao primeiro espectrômetro (espectrômetro de referência): um coeficiente de determinação ajustado (R^2_p) de aproximadamente 0,842, um erro padrão de previsão (RMSEP) de 0,017 e um índice de desempenho potencial (RPD) de 2,544. A técnica de ensemble stacking, aliada a métodos de transferência de calibração, conseguiu simplificar o processo de calibração cruzada entre os diferentes dispositivos, obtendo previsões comparáveis às de modelos tradicionais, porém com menos complexidade e maior eficiência.

Título Artigos	Ano	Nutriente	Qualis	JCR	Quantidade de Amostras	Algoritmos Utilizados	Região	Variáveis utilizadas	Resultados obtidos
Comparison of Soil Total Nitrogen Content Prediction Models Based on Vis-NIR Spectroscopy (Zheng et al., 2020)	2020		A1	SIM	19.036	Random Forest OLSE Extreme Learning Machine (ELM)	Países Europeus	Teor de nitrogênio total Outros parâmetros físicos e químicos do solo	O modelo com mais camadas de convolução (Model 3) alcançou um coeficiente de determinação (R^2) de aproximadamente 0,93, e um erro padrão de previsão absoluto médio (RMSEP) de 0,95 g/kg na previsão do STN, indicando excelente desempenho

Fonte: O autor.

Com base na análise comparativa dos artigos apresentados na Tabela 1, observa-se predominância de estudos voltados à predição de COT e NT, elementos centrais na avaliação da matéria orgânica e da fertilidade do solo. Esses dois nutrientes aparecem em maior número de trabalhos, seguidos por investigações sobre enxofre e boro, que, embora menos frequentes, demandam metodologias específicas devido às suas baixas concentrações e à dificuldade de quantificação precisa.

Entre as técnicas espectroscópicas empregadas, a espectroscopia no infravermelho próximo é a mais recorrente. A espectroscopia no infravermelho médio (MIR) é uma alternativa, também usada em combinação com NIR, e vários trabalhos demonstram melhor desempenho preditivo quando MIR é utilizada isoladamente ou integrada à NIR. Para nutrientes específicos, como enxofre e boro, destacam-se técnicas avançadas, como XANES (Espectroscopia de Absorção de Raios X de Energia) e WD-XRF (Espectroscopia de Fluorescência de Raios X por Dispersão de Comprimento de Onda), capazes de identificar formas químicas e fornecer informações mais detalhadas sobre a disponibilidade e a especiação desses elementos. A técnica LIBS (Espectroscopia de Emissão Óptica por Plasma Induzido por Laser) também aparece pontualmente, mostrando potencial para análise de enxofre em comprimentos de onda específicos da região VUV do espectro.

Quanto aos algoritmos e métodos de modelagem, o PLSR é amplamente empregado como padrão devido à sua robustez e à interpretação fácil. No entanto, estudos comparativos indicam que abordagens alternativas, como MARS (*Multivariate Adaptive Regression Splines*), podem superar o desempenho do PLSR e da SVM em determinadas condições, especialmente para propriedades específicas do solo. Nota-se também que, tratando-se de espectroscopia NIR, frequentemente são usados os algoritmos PLSR, SVM e, em alguns casos, ANNs.

A escala e o número de amostras variam significativamente entre os trabalhos, indo de conjuntos reduzidos de 10 a 50 amostras a bases de dados com mais de 19.000 amostras. Estudos de maior escala tendem a focar em calibrações generalistas e na aplicabilidade em diferentes contextos, enquanto pesquisas com menor número de amostras exploram metodologias mais específicas, testando abordagens inovadoras ou técnicas de maior resolução.

A diversidade geográfica também é um aspecto relevante, com trabalhos desenvolvidos em diferentes continentes, abrangendo desde regiões tropicais até áreas polares.

4 MATERIAL E MÉTODOS

4.1 Base de Dados

Neste estudo, utilizou-se uma abordagem baseada em espectroscopia NIR e AM para estimar os teores de COT e de N em amostras de solo.

A base de dados utilizada compreende 235 amostras. A coleta de solo foi realizada em uma área experimental localizada no Centro de Pesquisa para Assistência e Difusão Técnica Agropecuária da Fundação (ABC), no município de Ponta Grossa. A área localiza-se nas coordenadas 25º00’35"S e 50º09’16"W. Após a coleta, as amostras foram analisadas no Laboratório de Fertilidade do Solo da Universidade Estadual de Ponta Grossa (UEPG).

Para cada amostra, foram obtidos em laboratório os valores de referência de N e COT. Os atributos N e COT foram utilizados separadamente ao longo de todo o estudo.

As leituras espectrais de todas as amostras foram previamente obtidas no equipamento NIRS DA1650, disponível no Complexo de Laboratórios Multiusuários da UEPG (CLABMU), na faixa espectral de 1.100 nm a 1.648 nm, com intervalo de leitura de 2 nm, totalizando 277 atributos por amostra.

Dessa forma, cada amostra da base de dados contém: teor de COT, teor de N e mais 277 atributos, sendo cada um deles a resposta espectral obtida para a amostra no respectivo comprimento de onda.

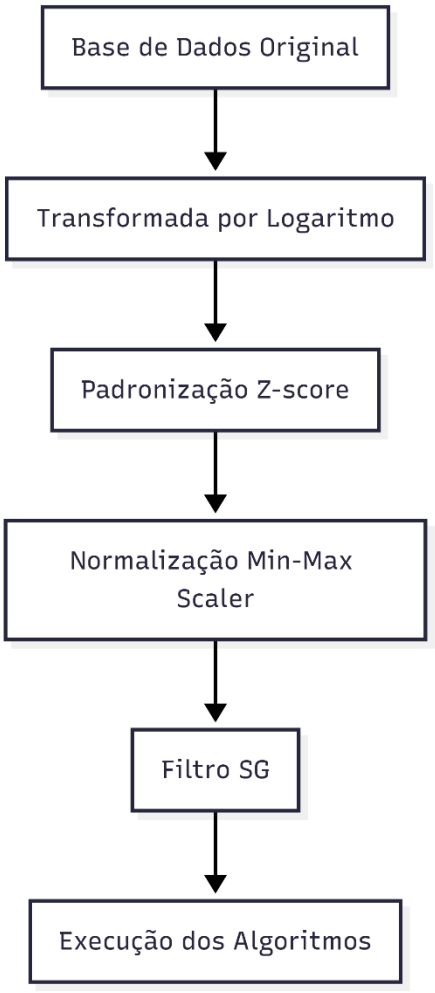
4.2 Ambiente Computacional

Os experimentos foram conduzidos utilizando a linguagem de programação Python, por sua ampla adoção em ciência de dados e aprendizado de máquina, bem como pela disponibilidade de bibliotecas consolidadas para análise espectral, pré-processamento e modelagem estatística. Todo o desenvolvimento foi realizado no ambiente Google Colab, que ofereceu uma infraestrutura computacional estável e reprodutível, incluindo acesso facilitado a GPU/CPU. Esse ambiente permitiu executar todas as etapas do fluxo de trabalho, desde o carregamento e a organização dos dados espectrais, a aplicação dos métodos de pré-processamento, a seleção de variáveis e o treinamento dos modelos, até a validação cruzada e a análise dos resultados de forma eficiente e padronizada.

4.3 Pré-processamento

No pré-processamento, adotou-se uma sequência estruturada de transformações aplicadas à base de dados, com o objetivo de investigar o impacto de cada etapa nos modelos preditivos. A Figura 8 abaixo apresenta essa sequência, iniciando nos dados originais e avançando pelas operações de transformação logarítmica, padronização, normalização e aplicação do filtro de Savitzky–Golay (SG), até a etapa final de execução dos algoritmos.

Figura 8. Fluxo de Pré-processamento dos Dados



Fonte: O autor.

A partir da base de dados original, aplicou-se a transformação logarítmica para reduzir a assimetria das variáveis e suavizar a influência de valores extremos, melhorando a linearidade entre os atributos de comprimento de onda e os atributos N e COT.

Em seguida, foram realizadas a padronização dos dados e a minimização dos valores anômalos, utilizando o método Z-*score*.

Posteriormente, os dados foram normalizados com o algoritmo MinMaxScaler, ajustando todas as variáveis para o intervalo de 0 a 1, evitando que diferenças de escala prejudicassem os modelos.

Para a suavização espectral, aplicou-se o filtro de Savitzky-Golay, que reduz o ruído sem alterar significativamente a forma e os picos do sinal.

Inicialmente, foram utilizados os algoritmos de AM com a base de dados contendo todos os 235 atributos. Em seguida, considerando a alta dimensionalidade da base de dados, foi aplicado o método de seleção de atributos SPA, em conjunto com os algoritmos *Linear Regression e* PLSR.

4.4 Algoritmos Usados Para a Criação dos Modelos de Predição

Foram empregados diferentes algoritmos de regressão, abrangendo distintas abordagens estatísticas e de AM, sendo eles: *Linear Regression*, *PLSR, MLPRegressor, Random Forest Regressor, SVR (Support Vector Regression) e OMP (Orthogonal Matching Pursuit).*

A validação dos modelos foi conduzida por meio de validação cruzada 10-*Folds*, garantindo maior robustez e confiabilidade na avaliação do desempenho preditivo (KUHN e JOHNSON, 2013). Para isso, o conjunto de dados foi particionado em dez subconjuntos de tamanho aproximadamente igual. Em cada iteração, nove *folds* foram utilizados para treinamento, enquanto o restante foi destinado ao teste. Esse processo foi repetido até que cada subconjunto fosse utilizado exatamente uma vez como conjunto de validação.

As métricas finais de desempenho foram calculadas a partir da média dos resultados obtidos nos dez *folds*, assegurando uma avaliação mais robusta, estável e confiável, uma vez que todos os dados desempenham, em momentos distintos, os papéis de treino e teste.

A avaliação dos modelos gerados pelos algoritmos foi realizada por meio das métricas: coeficiente de determinação (R^2), erro absoluto médio (MAE, do inglês *Mean Absolute Error*) e o erro percentual absoluto médio (MAPE, do inglês *Mean Absolute Percentual Error)*.

4.5 Parametrização dos Algoritmos

A parametrização dos algoritmos constitui uma etapa fundamental no desenvolvimento de modelos preditivos, uma vez que influencia diretamente a capacidade de generalização, a estabilidade e o desempenho das abordagens empregadas.

Este trabalho, buscou-se adotar configurações que equilibrassem complexidade e robustez, considerando as características dos dados espectrais. Nas seções 4.5.1 a 4.5.9, são descritos detalhadamente os procedimentos adotados para a implementação de cada algoritmo para predição de N e COT.

4.5.1 Rede Neural Artificial – MLP

O modelo baseado em Perceptron Multicamadas (MLP) foi implementado utilizando o algoritmo *MLPRegressor*, com arquitetura profunda composta por três camadas ocultas contendo 300, 150 e 60 neurônios, respectivamente (*hidden_layer_sizes = (300, 150, 60)*).

A função de ativação utilizada foi a ReLU (*activation = 'relu'*), e o treinamento foi conduzido com o otimizador Adam (*solver = 'adam'*), com taxa de aprendizado inicial de 0,0002 (*learning_rate_init = 0.0003*). Para controle de regularização, foi adotado o parâmetro *alpha = 0.0005*, reduzindo o risco de overfitting.

O treinamento foi limitado a 4000 iterações (*max_iter = 4000*) e incorporou o mecanismo de *early stopping*, com 15% dos dados reservados para validação (*validation_fraction = 0.15*) e critério de parada baseado em 30 iterações consecutivas sem melhoria (*n_iter_no_change = 30*).

4.5.2 Regressão Linear

O modelo de regressão linear foi implementado utilizando o método dos mínimos quadrados ordinários por meio da classe *LinearRegression.*

Não foram aplicados parâmetros de regularização, mantendo-se a formulação clássica da regressão linear. O modelo foi ajustado diretamente aos dados de treinamento e avaliado no conjunto de teste em cada fold.

4.5.3 Regressão *Ridge*

A regressão *Ridge* foi empregada com o objetivo de mitigar problemas de multicolinearidade, utilizando penalização L2. O modelo foi implementado por meio da classe *Ridge*, com parâmetro de regularização fixado em *alpha = 1.0*.

4.5.4 PLSR

O modelo PLSR também foi aplicado sem seleção prévia de variáveis, utilizando todas as variáveis espectrais disponíveis. O número de componentes latentes foi definido como 35 (*n_components = 35*), visando capturar maior variabilidade dos dados.

Além das métricas tradicionais, foi realizada a análise de importância das variáveis por meio do cálculo do (VIP), permitindo identificar os comprimentos de onda mais relevantes para a predição.

4.5.5 PLSR com seleção de variáveis (SPA-PLSR)

Para redução da dimensionalidade espectral, foi aplicada a técnica de seleção de variáveis (SPA), previamente à modelagem. Foram selecionadas 75 variáveis espectrais (*n_vars = 75*), reduzindo redundância e colinearidade.

O modelo de regressão foi construído utilizando o método dos mínimos quadrados parciais por meio da classe *PLSRegression*, com 15 componentes latentes (*n_components = 15*).

4.5.6 Regressão Linear com SPA

Uma abordagem híbrida foi adotada combinando o algoritmo de seleção de variáveis (SPA) com a regressão linear clássica implementada via *LinearRegression*.

Inicialmente, o SPA foi aplicado com o objetivo de reduzir a dimensionalidade do conjunto espectral, selecionando 75 variáveis (*num_vars = 75*). Esse valor foi definido buscando um equilíbrio entre redução de redundância e preservação da informação relevante, uma vez que dados espectrais apresentam alta colinearidade entre comprimentos de onda adjacentes.

4.5.7 *Random Forest*

O modelo baseado em florestas aleatórias foi implementado por meio da classe *RandomForestRegressor*, com n_estimators = 250, ou seja, um conjunto de 250 árvores de decisão.

Cada árvore foi construída a partir de subconjuntos aleatórios dos dados e das variáveis, promovendo diversidade entre os estimadores.

O parâmetro *random_state* foi definido dinamicamente com base no número do *fold* durante a validação cruzada, garantindo reprodutibilidade controlada e, ao mesmo tempo, introduzindo variabilidade entre os modelos treinados em cada partição. Os demais parâmetros foram mantidos como padrão da biblioteca, incluindo *max_depth = None* (crescimento completo das árvores), *min_samples_split = 2* (isso significa que qualquer nó que tenha pelo menos 2 amostras pode ser dividido) e *min_samples_leaf = 1* (significando que uma folha pode conter apenas uma única amostra).

Essa configuração permite que o modelo capture relações não lineares e interações entre variáveis, sendo particularmente útil em dados espectrais com comportamento complexo.

4.5.8 *Support Vector Regression* (SVR)

O modelo de regressão por vetores de suporte foi implementado utilizando a classe SVR, com kernel radial (*kernel = 'rbf'*), amplamente utilizado para modelagem de relações não lineares.

O kernel RBF projeta os dados para um espaço de maior dimensionalidade, permitindo a separação de padrões complexos. Os principais parâmetros utilizados foram mantidos como padrão, *C = 1.0* controla o trade-off entre erro de treinamento e complexidade do modelo, *epsilon = 0.1* define a margem de tolerância para erros e *gamma = 'scale'* ajusta automaticamente a influência de cada ponto de treinamento.

A escolha de manter os parâmetros padrão está associada à robustez inicial do modelo, embora o SVR seja conhecido por sua alta sensibilidade à parametrização.

4.5.9 *Orthogonal Matching Pursuit* (OMP)

O algoritmo de seleção esparsa foi implementado por meio da classe *OrthogonalMatchingPursuit*, sendo uma abordagem baseada na decomposição do sinal em um conjunto reduzido de variáveis.

Neste trabalho, não foi definido explicitamente o número de coeficientes não nulos (*n_nonzero_coefs*), permitindo que o modelo determinasse automaticamente o nível de esparsidade com base nos dados.

5 RESULTADOS E DISCUSSÃO

Os algoritmos foram executados com a base de dados original e com várias estratégias de transformação de dados, como a transformada por logaritmo, a padronização (Z-score), a normalização (Min-Max Scaler) e a padronização com o filtro SG. Os resultados apresentados nessa seção referem-se às melhores estratégias obtidas.

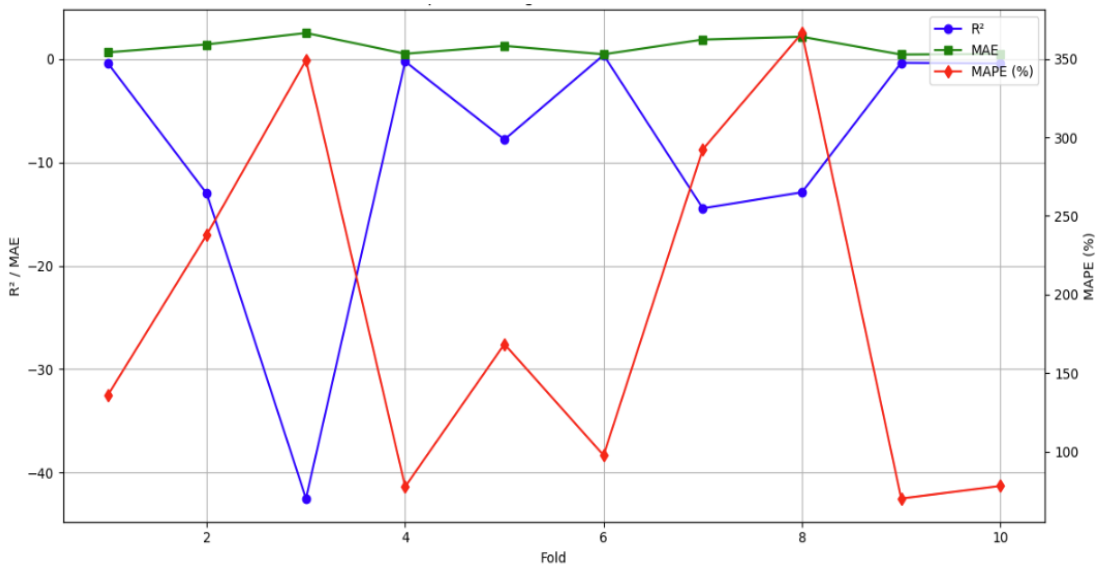
5.1 Resultados da Execução dos Algoritmos para a Estimativa de N

Para a predição do teor de N, foram empregados os seguintes algoritmos de aprendizado de máquina: *Linear Regression*, PLSR combinado com SPA, MLP *Regressor*, *Linear Regression + SPA*, SVR, *Random Forest Regressor* e OMP. O desempenho de cada modelo foi avaliado por meio do método de validação cruzada K-Fold com k = 10, e todas as métricas (R^2, MAE e MAPE) foram calculadas como a média dos 10 *folds* de cada algoritmo.

5.1.1 *Linear Regression*

No caso da Regressão Linear aplicada ao nitrogênio, por meio do algoritmo *Linear Regression,* o R^2 médio foi de -9,1602, indicando que o modelo não foi capaz de explicar a variabilidade dos dados, uma vez que um R^2 negativo evidência que o modelo não acompanha a tendência observada, ajustando-se de forma inferior a uma linha horizontal, conforme ilustrado na Figura 9.

Figura 9. Desempenho da *Linear Regression* na estimativa de N.



Fonte: O autor.

Isso sugere que a relação entre os preditores espectrais e o nitrogênio não é adequadamente capturada por uma regressão linear simples.

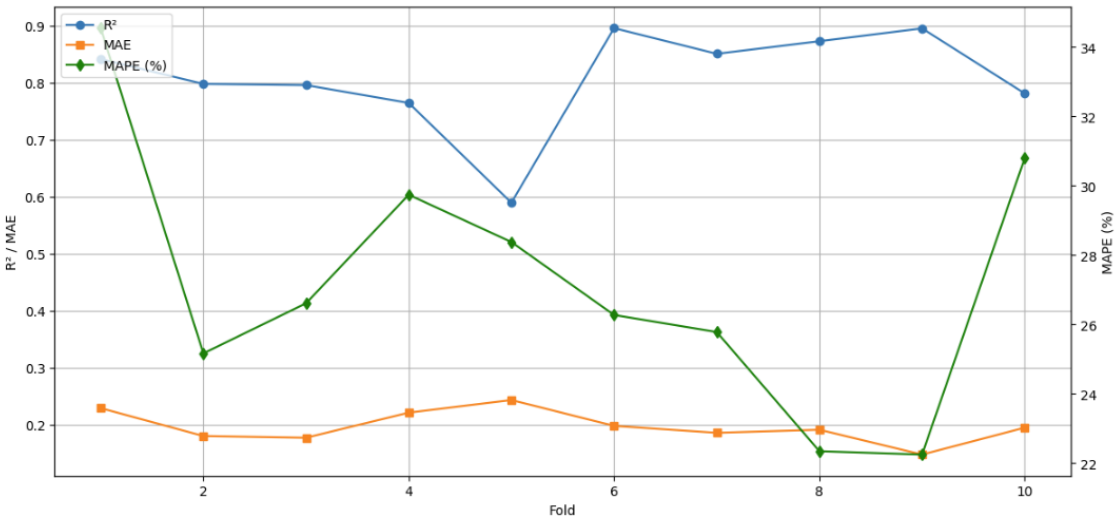
As métricas de erro confirmam essa inadequação, com MAE médio de 1,1901 e MAPE médio de 187,45%, revelando erros absolutos elevados e uma incapacidade significativa de generalização.

5.1.2 PLS *Regression* + SPA

O modelo *PLS Regression*, combinado com a técnica de seleção de variáveis SPA, apresentou desempenho satisfatório na predição da variável de interesse. O modelo usou as 75 faixas espectrais que apresentaram maior correlação com o atributo N e teve o R^2 médio de 0,8082, MAE de 0,1962 e MAPE de 27,18%.

Esses resultados mostram que o uso da seleção de variáveis com SPA contribuiu para reduzir a complexidade do modelo, mantendo uma boa precisão preditiva. Detalhes dos resultados de cada execução estão apresentados na Figura 10.

Figura 10. Desempenho do PLSR + SPA para estimativa de N

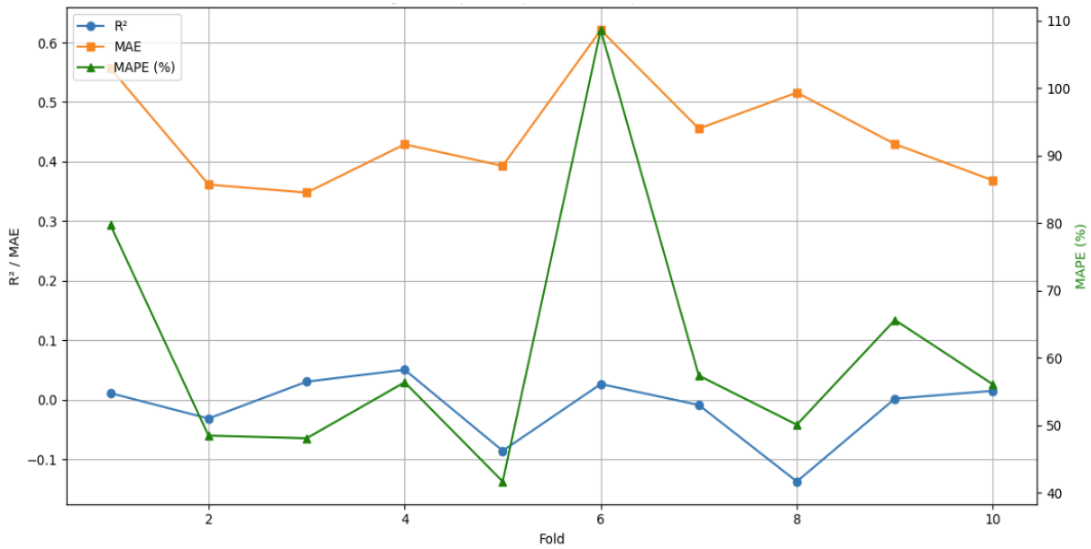


Fonte: O autor.

5.1.3 MLP *Regressor*

No caso do algoritmo MLP *Regressor*, o MAPE foi de 61,21% e o MAE médio de 0,4479, o que indica que, em termos absolutos, o erro foi menor do que o da Linear Regression, mas ainda assim insatisfatório (Figura 11).

Figura 11. Valores de R², MAE e MAPE do algoritmo MLP *Regressor* para estimativa de N.



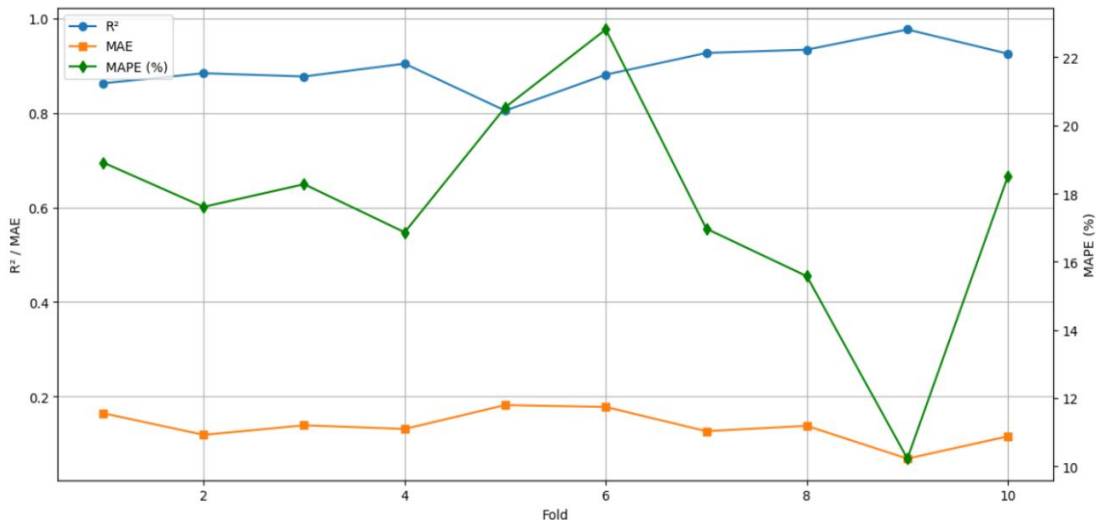
Fonte: O autor.

Por outro lado, o R^2 médio de -0,0129 evidencia que o modelo não conseguiu explicar adequadamente a variabilidade do nitrogênio e, sendo ainda negativo, demonstra que o desempenho do MLP foi inferior a uma simples predição pela média dos valores observados.

5.1.4 *Linear Regression + SPA*

O algoritmo apresentou um valor médio de R^2 de 0,8975, indicando que explicou uma grande parte da variabilidade do nitrogênio a partir dos atributos espectrais. O valor é considerado satisfatório, representando um avanço em relação à Regressão Linear simples sem seleção de atributos, que apresentou R^2 negativo. O MAE foi de 0,1365 e o MAPE de 17,63%. A Figura 12 apresenta o gráfico de execuções do algoritmo para estimativa de N.

Figura 12. Desempenho algoritmo de *Linear Regression + SPA* para estimativa de N.

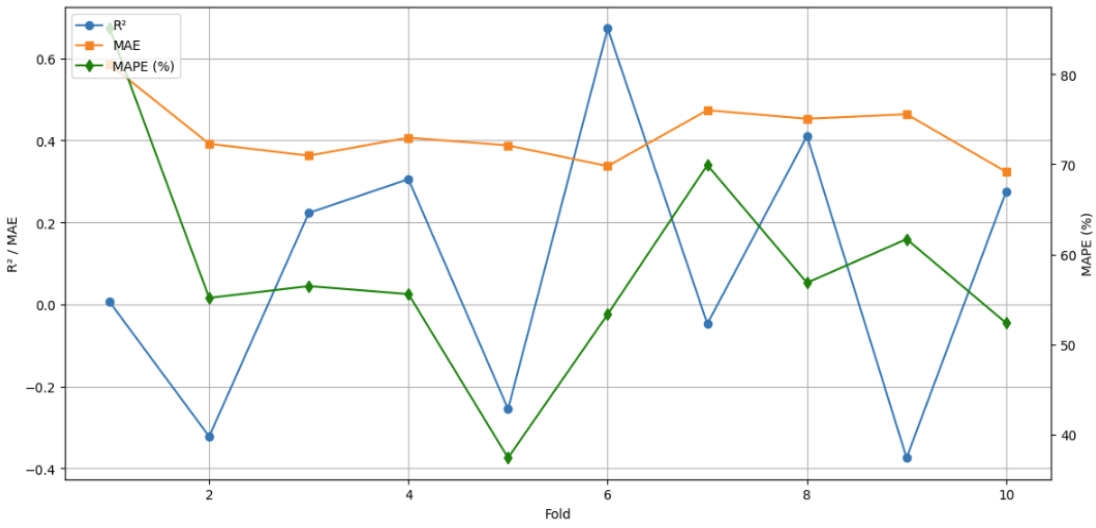


Fonte: O autor.

5.1.5 *Random Forest Regressor*

O algoritmo *Random Forest Regressor* foi aplicado à predição do teor de N, também com validação cruzada, e avaliado pela média das métricas. Na Figura 13 está apresentado o desempenho do algoritmo ao longo da validação cruzada.

Figura 13. Desempenho do algoritmo *Random Forest Regressor* na estimativa de N.



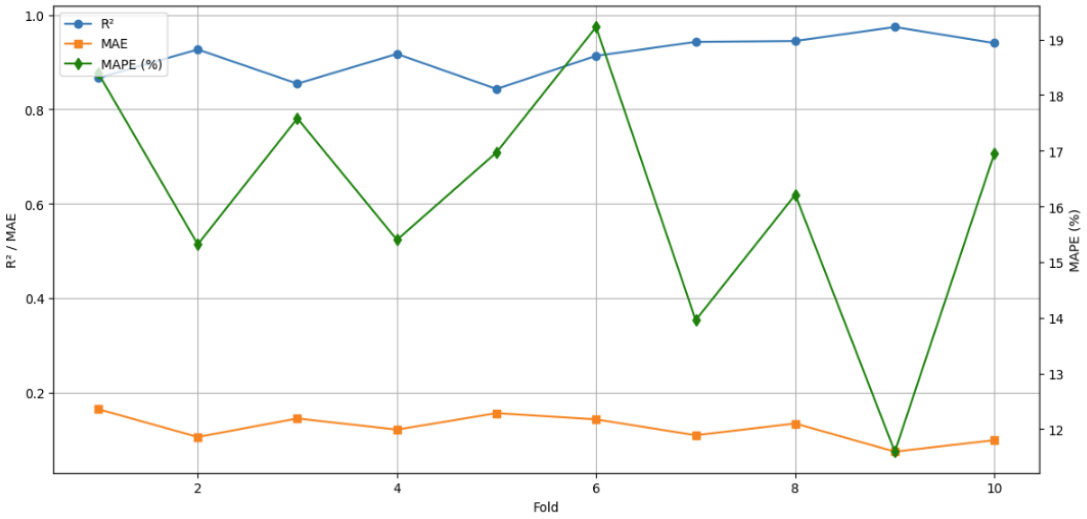
Fonte: O autor.

O valor médio de R^2 desse algoritmo foi de 0,0896, indicando que o modelo não conseguiu explicar a variabilidade dos dados, assim como os algoritmos previamente utilizados (seção 5.1.1).

5.1.6 PLS *Regression*

Na predição do teor de N, o modelo gerado com o algoritmo PLS *Regression* apresentou um dos melhores desempenhos, entre todos os algoritmos testados. Após a execução, obteve média de R^2 de 0,9124, MAE de 0,1253 e MAPE de 16,16%, indicando que conseguiu explicar boa parte da variabilidade dos dados ao correlacionar os espectros NIR com o teor de nitrogênio, como pode ser observado na Figura 14.

Figura 14. Desempenho do algoritmo PLSR para estimativa de N.

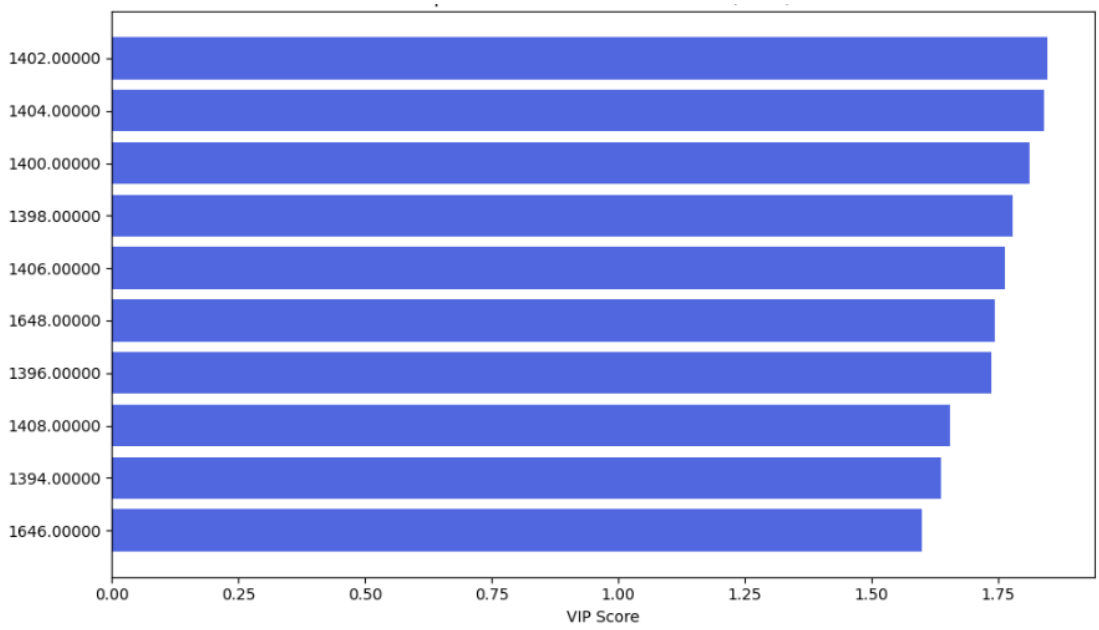


Fonte: O autor.

Com o objetivo de identificar os atributos preditores mais relevantes para N, foi aplicado o modelo PLSR, em conjunto com a métrica VIP (do inglês, *Variable Importance in Projection*), usada em modelos de PLSR para avaliar a importância de cada variável explicativa na predição da variável resposta. Ela indica quais atributos contribuem mais para explicar a variabilidade do modelo. Em geral, variáveis com VIP maior que 1 são consideradas importantes, enquanto as com VIP menor que 0,8 têm pouca relevância. Valores entre 0,8 e 1 indicam relevância moderada. Essa métrica é útil em dados espectrais, porque permite identificar atributos (comprimentos de onda) mais significativos para a predição, facilitando a interpretação do modelo.

A análise da VIP revelou os 10 atributos mais importantes, cujos valores de VIP indicam uma grande contribuição para o modelo. Os comprimentos de onda mais significativos são apresentados na Figura 15.

Figura 15. Os 10 Atributos mais relevantes para a estimativa de NT



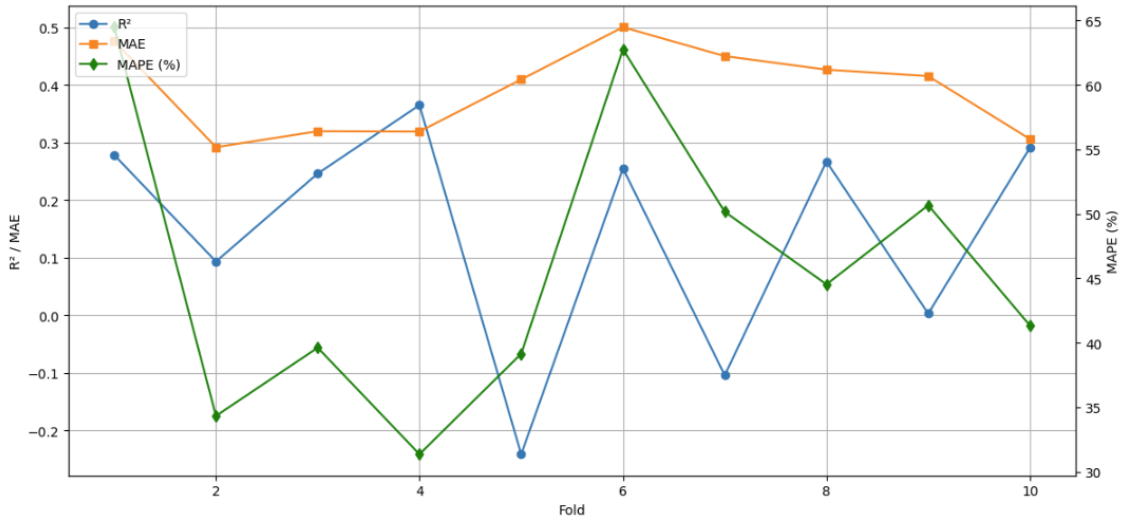
Fonte: O autor.

Todos os valores de VIP apresentados no gráfico da Figura 14 estão acima de 1, o que evidencia que essas faixas espectrais são altamente relevantes para a predição do nitrogênio e podem ser priorizadas em futuras análises ou na simplificação do modelo. Desta forma, os 10 comprimentos de onda mais importantes são 1402, 1404, 1400, 1398, 1406, 1648, 1396, 1408, 1394 e 1646 nm.

5.1.7 SVR

Os modelos gerados pelo algoritmo SVR resultaram em um R^2 médio baixo, de 0,1454, mostrando-se incapazes de estimar N a partir dos valores das bandas espectrais contidas na base de dados, como apresentado no gráfico da Figura 16.

Figura 16. Valores de R², MAE e MAPE do algoritmo SVR para estimativa de N.

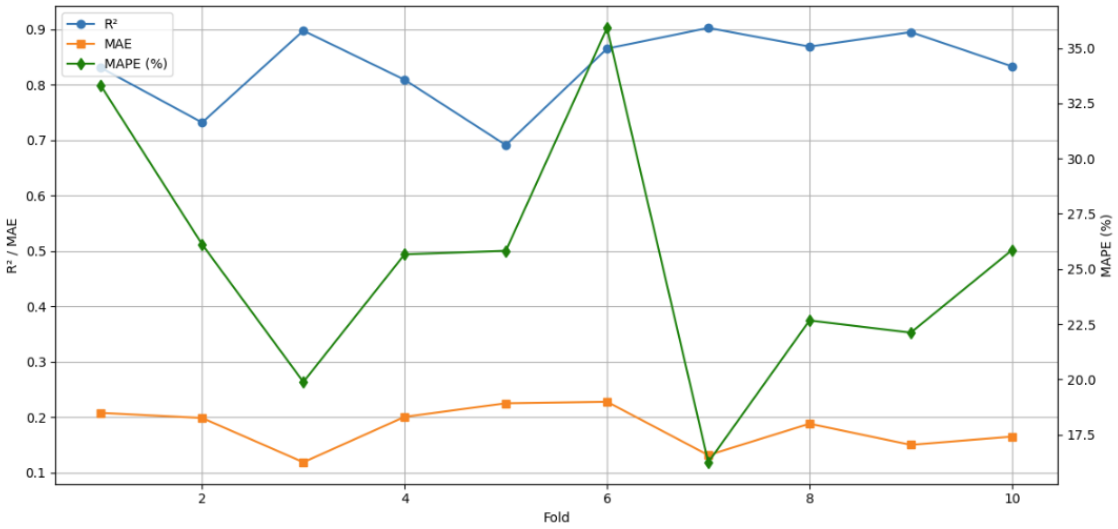


Fonte: O autor.

5.1.8 OMP

Na predição do teor de N, o algoritmo OMP apresentou excelente desempenho. Após a execução, o modelo obteve R² médio de 0.8327, MAE de 0.1816 e MAPE 25.36%, indicando que conseguiu explicar a maior parte da variabilidade dos dados e gerar previsões próximas dos valores reais observados. Esses resultados mostram que o OMP foi eficiente na captura das relações entre os atributos espectrais e o teor de nitrogênio total Figura 17.

Figura 17. Desempenho do algoritmo OMP para estimativa de N.



Fonte: O autor.

O Quadro 1 contém uma síntese dos resultados obtidos para estimativa de N

Quadro 1. Média das métricas de R^2, MAE e MAPE (%) para algoritmos de predição de Nitrogênio.

Algoritmos	**R^2**	**MAE**	**MAPE (%)**
PLSRegression	0,9124	0,1253	16,16
LinearRegression + SPA	0,8975	0,1365	17,63
OrthogonalMatchingPursuit	0,8327	0,1816	25,36
PLSR + SPA	0,8082	0,1962	27,18
SVR	0,1454	0,3916	45,84
RandomForestRegressor	0,0896	0,4184	58,39
MLPRegressor	-0,0129	0,4479	61,21
LinearRegression	-9,1602	1,1901	187,45

Fonte: O autor.

5.1.9 Desempenho na predição de N

O desempenho obtido para o modelo PLRS, com R^2 de validação igual a 0,9124, apresenta um posicionamento competitivo quando comparado a diferentes abordagens de ponta relatadas na literatura recente. Em relação ao modelo de Liu et al. (2025), que utilizou o conceito de SVM híbrido (RBF-poly-SVM) otimizado pelo Algoritmo de Otimização de Baleias (WOA)

e associado à seleção de variáveis via IRIV, esse resultado, utilizando PLRS, teve desempenho um pouco superior (0,9124), já que o modelo de Liu et al. (2025) apresentou R^2 de 0,902, mesmo sendo um método de aprendizagem de máquina altamente otimizado.

Zheng et al. (2020) compararam modelos de previsão de N baseados em espectroscopia Vis-NIR, utilizando aprendizado de máquina convencional e aprendizado profundo. Utilizaram três métodos convencionais de AM, incluindo a estimativa por mínimos quadrados ordinários (OLSE), floresta aleatória (RF) e a máquina de aprendizado extremo (ELM), e também o método de aprendizado profundo com três estruturas diferentes de rede neural convolucional (CNN). Ao comparar os resultados de Zheng et al. (2020) com os obtidos nesse trabalho por meio de PLSR, vê-se que o modelo PLRS é inferior ao modelo baseado em *Deep Learning*, que atingiu R^2 de 0,93 em um grande conjunto de dados com mais de 19 mil amostras. Contudo, o resultado mostra-se superior ao do melhor modelo de AM convencional reportado pelos mesmos autores, baseado em ELM com correção de linha de base e suavização, cujo R^2 foi de 0,89.

Quando comparado ao estudo de Santos et al. (2020), que avaliou a técnica PLSR na região do infravermelho médio (MIR) e obteve R^2 de 0,90, o modelo PLSR apresenta desempenho altamente comparável, sendo significativo notar que a região MIR costuma oferecer maior acurácia nas predições de nitrogênio do que abordagens baseadas apenas no espectro Vis-NIR. Assim, alcançar um R^2 superior ao obtido no PLSR em MIR evidencia a robustez da estratégia empregada.

Em relação ao trabalho de Tao et al. (2025), que utilizou um *ensemble stacking* envolvendo PLSR, SVR, *Ridge e BoostForest* aplicado à transferência de calibração entre instrumentos, o PLSR apresenta um resultado claramente superior em seu próprio contexto experimental, visto que o *ensemble* atingiu R^2 de 0,842.

Quadro 2. Comparação trabalhos correlatos para estimativa de Nitrogênio Total

Modelo de Referência	**Estratégia Principal**	**R^2 de (N)**	**Comparação com resultado PLRS (R^2: 0,9124)**
(Liu et al., 2025)	SVM Híbrido (RBF-poly-SVM) otimizado por Algoritmo de Otimização de Baleias (WOA) + Seleção de Variáveis (IRIV)	0,902	O resultado é superior ao deste modelo de Aprendizado de Máquina altamente otimizado.
(Zheng et al., 2020)	CNN (Deep Learning, Modelo 3) em grande *dataset* (19.019 amostras)	0,93	O resultado é inferior ao do modelo de Deep Learning mais avançado, mas superior ao do melhor modelo de AM convencional (ELM, $R^2 = 0,89$).
(Zheng et al., 2020)	AM convencional (ELM) + Correção de Linha de Base/Suavização	0,89	O resultado é superior ao do melhor modelo de AM convencional, o ELM, em grandes *datasets*.
(Santos et al., 2020)	PLSR em espectroscopia no infravermelho médio (MIR)	0,90	O resultado é altamente comparável ao melhor desempenho obtido com a região MIR, que tipicamente oferece maior acurácia em N do que apenas o Vis-NIR.
(Tao et al., 2025)	*Ensemble Stacking* (PLSR, SVR, Ridge + BoostForest) para *Calibration Transfer* (transferência entre instrumentos)	0,842	O resultado é superior ao desempenho de transferência interinstrumental, que, naturalmente, apresenta desafios e degradação de desempenho.

Fonte: O autor.

5.2 Resultados da Execução dos Algoritmos para Estimativa de COT

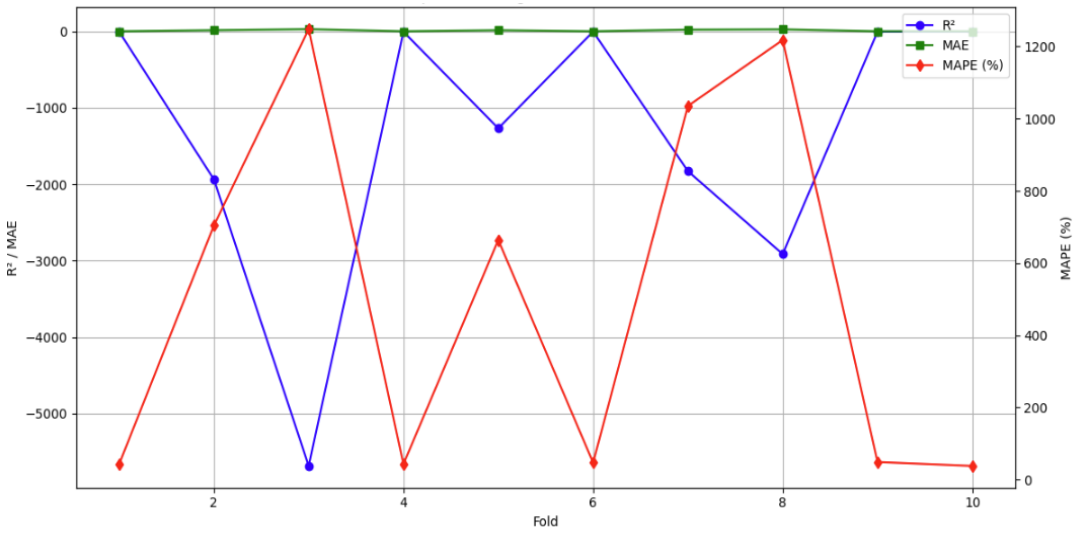
Os modelos para a predição do teor de COT foram gerados com os mesmos algoritmos empregados na estimativa de N, a saber: Linear Regression, Linear Regression + SPA, PLS Regression, PLSR + SPA, MLP *Regressor*, *Random Forest Regressor*, SVR e OMP. Além destes, para a estimativa de COT, utilizou-se também o algoritmo *Ridge Regression.*

Também foi realizada a associação dos algoritmos Linear Regression e PLSR ao método de seleção de atributos SPA.

5.2.1 *Linear Regression*

O algoritmo de *Linear Regression* apresentou resultados insatisfatórios na estimativa de COT. O R^2 médio foi de -1363,8743, o MAE foi de 11,4907 e o MAPE foi de 508,80%, indicando desempenho drasticamente inferior ao de uma simples média dos valores observados, o que evidencia que o modelo não conseguiu capturar a relação entre as variáveis independentes e o COT. Esses resultados sugerem que a regressão linear múltipla, em sua forma simples, não foi adequada para modelar os dados espectrais relacionados ao COT, provavelmente devido à presença de multicolinearidade, não linearidade na relação entre variáveis ou à alta dimensionalidade dos dados. A Figura 18 apresenta as médias das execuções.

Figura 18. Desempenho da Regressão Linear para estimativa de COT

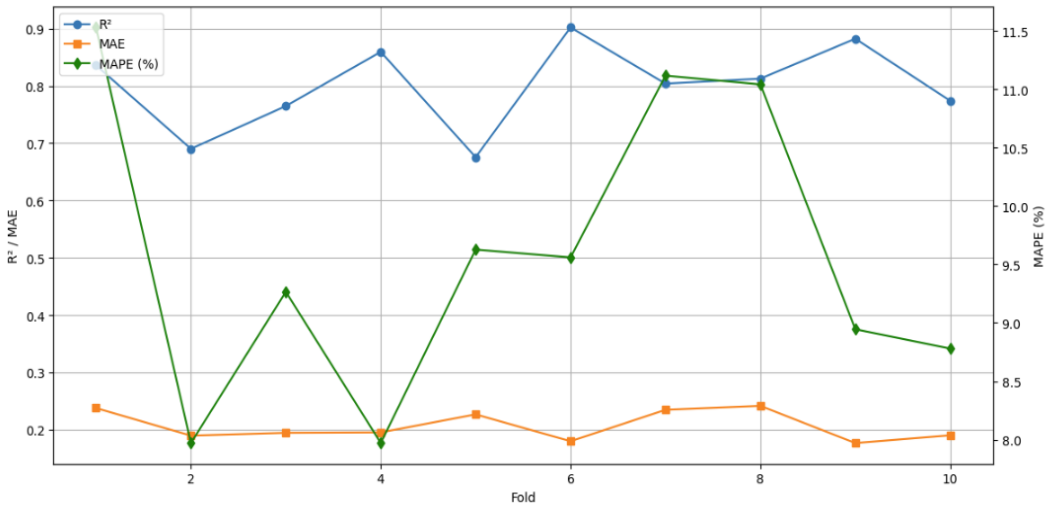


Fonte: O autor.

5.2.2 *Linear Regression* + SPA

Da mesma forma que a associação do algoritmo *Linear Regression* ao SPA melhorou a estimativa de N, essa associação também contribuiu para a melhora da estimativa do COT. A média do R^2 foi de 0,8003, o MAE foi de 0,2064 e o MAPE foi de 9,58%. Na Figura 19 estão apresentados os resultados das execuções ao longo dos 10 folds da validação cruzada.

Figura 19. Desempenho da *Linear Regression* + SPA para estimativa de COT



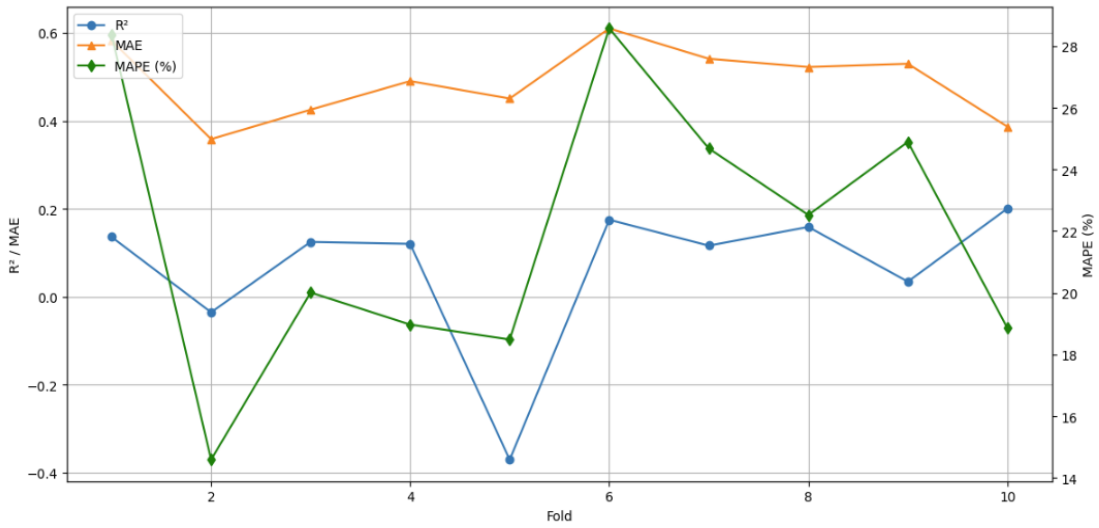
Fonte: O autor.

Os mesmos 75 atributos selecionados e usados para N também foram usados para COT, visto que a seleção de atributos pelo método SPA independe do atributo meta. Os atributos selecionados abrangem bandas espectrais de praticamente todas as regiões do espectro NIR utilizado neste estudo.

5.2.3 *Ridge Regression*

O algoritmo *Ridge Regression* apresentou desempenho insatisfatório na predição do COT. O R^2 médio foi de 0,0660, o MAE de 0,4895 e o MAPE de 22,00%. A Figura 20 contém o desempenho do modelo.

Figura 20. Desempenho do Ridge Regression para estimativa de COT

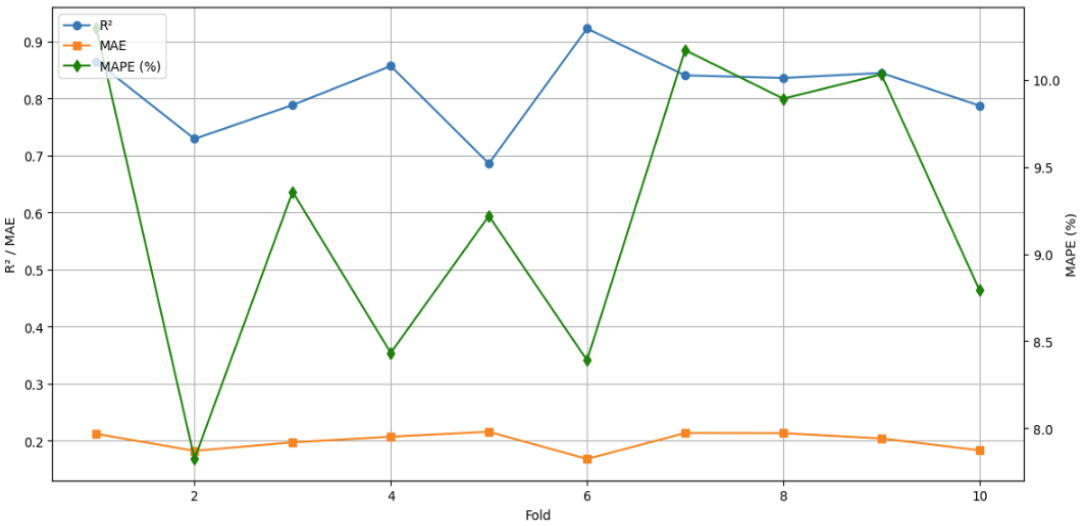


Fonte: O autor.

5.2.4 PLS *Regression*

O modelo PLSR apresentou um dos melhores desempenhos entre os algoritmos testados na predição do COT. O R^2 médio de 30 execuções foi de 0,8156, indicando que aproximadamente 81% da variabilidade de COT foi explicada pelos modelos gerados. Esse valor demonstra boa capacidade preditiva e representa um avanço substancial em relação aos demais algoritmos avaliados. O MAE foi de 0,1994, indicando que, em média, as previsões diferiram menos dos valores reais, e o MAPE médio foi de 9,24%. Os resultados confirmam a adequação do PLSR a dados espectrais, visto que o método é projetado para lidar com alta dimensionalidade e com forte correlação entre os atributos de predição, características típicas da espectroscopia. O gráfico do desempenho está apresentado na Figura 21.

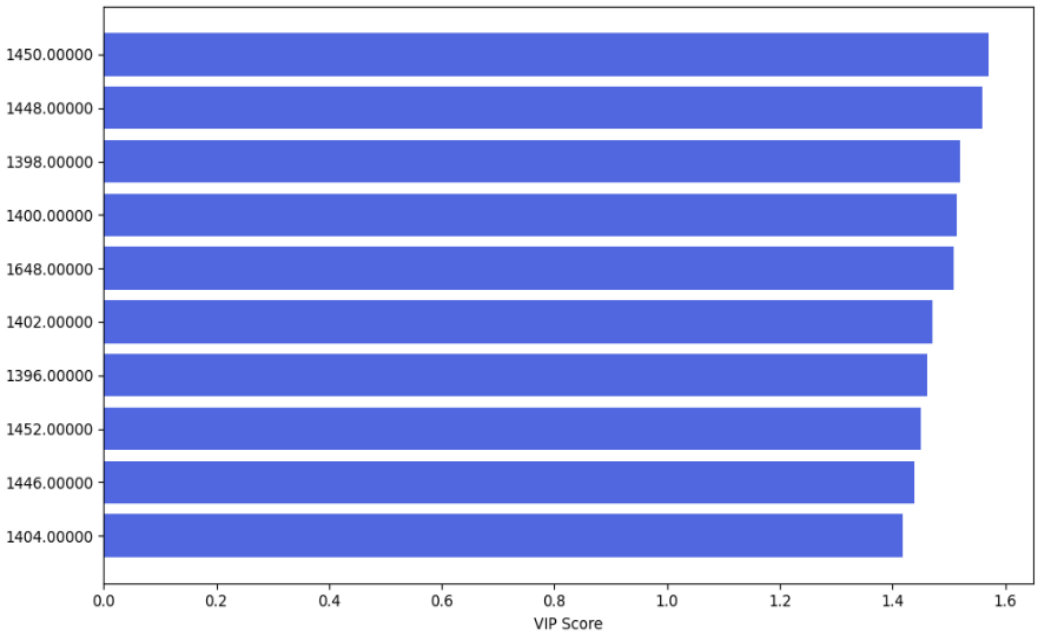
Figura 21. Desempenho do PLSR para estimativa de COT



Fonte: O autor.

Os dez atributos mais importantes identificados pelo critério VIP no modelo PLSR para estimativa de COT correspondem aos comprimentos de onda 1450, 1448, 1398, 1400, 1648, 1402, 1396, 1452, 1446 e 1404, com valores de VIP variando de 1,4 a 1,6, conforme apresentado na Figura 22.

Figura 22. Os 10 atributos mais relevantes para estimativa do COT

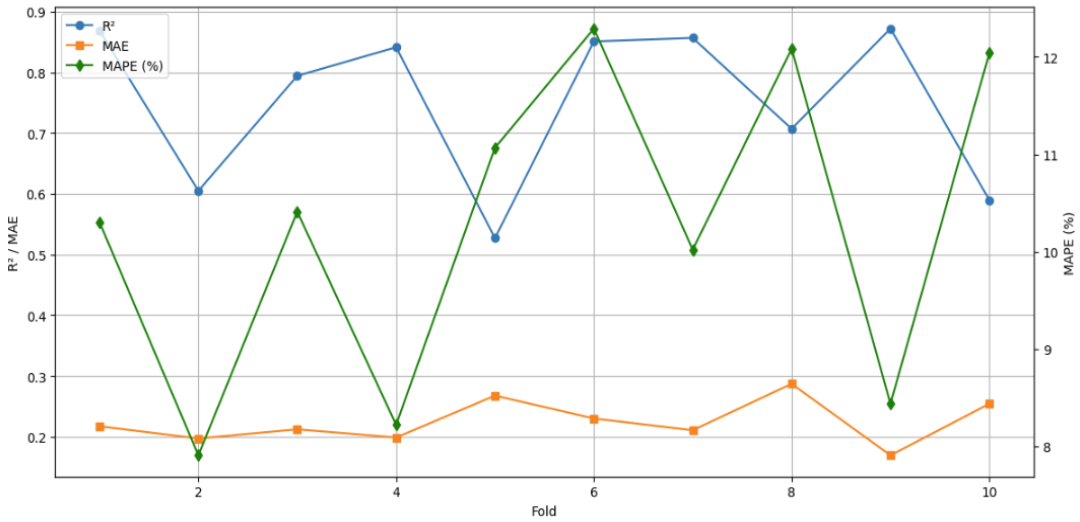


Fonte: O autor.

5.2.5 PLSR + SPA

O modelo PLSR combinado com SPA apresentou um desempenho consistente na análise dos dados. Os melhores resultados foram obtidos com a configuração do SPA para a seleção de 75 atributos. A média do R^2 foi de 0,7512, o MAE médio de 0,2246 e o MAPE de 10,28%. Esses resultados sugerem que a combinação de PLSR com SPA é uma boa alternativa para a estimativa de COT com base na faixa espectral estudada. A Figura 23 contém o gráfico do desempenho.

Figura 23. Valores de R^2, MAE e MAPE na validação cruzada de modelo gerado pelo algoritmo PLSR+ SPA para estimativa de COT

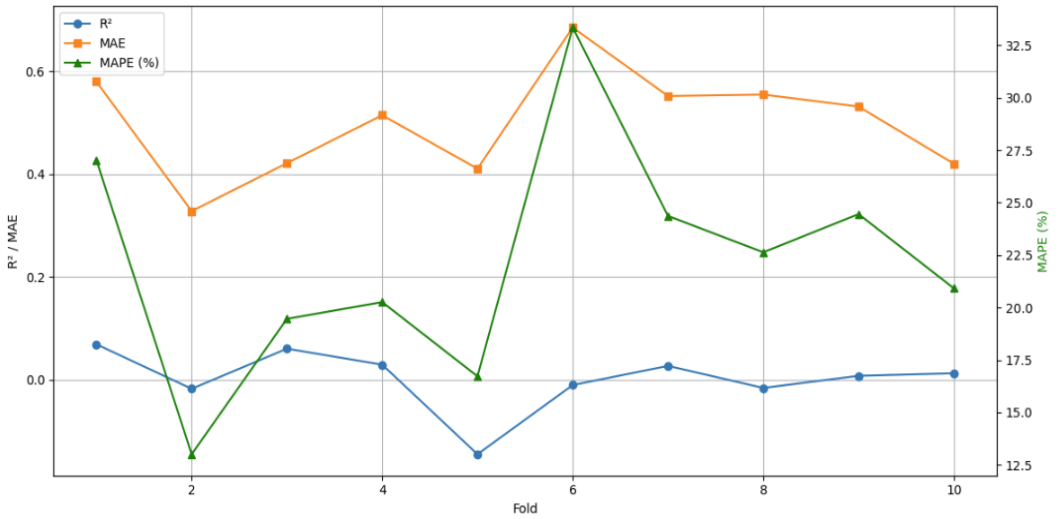


Fonte: O autor.

5.2.6 MLP *Regressor*

O algoritmo MLP *Regressor* apresentou desempenho insatisfatório na predição do teor de carbono. O R^2 médio foi 0,0019, o que indica que os modelos gerados pelo algoritmo não conseguiram explicar a variabilidade dos dados, performando pior do que uma simples média dos valores observados. O modelo também apresentou um MAE de 0,4996 e um MAPE de 22,22%. A Figura 24 apresenta o gráfico com os valores de R^2, MAE e MAPE obtidos nas execuções, indicando que a rede neural utilizada não foi adequada para identificar uma relação entre os atributos espectrais e o COT.

Figura 24. Desempenho MLP Regressor para estimativa de COT

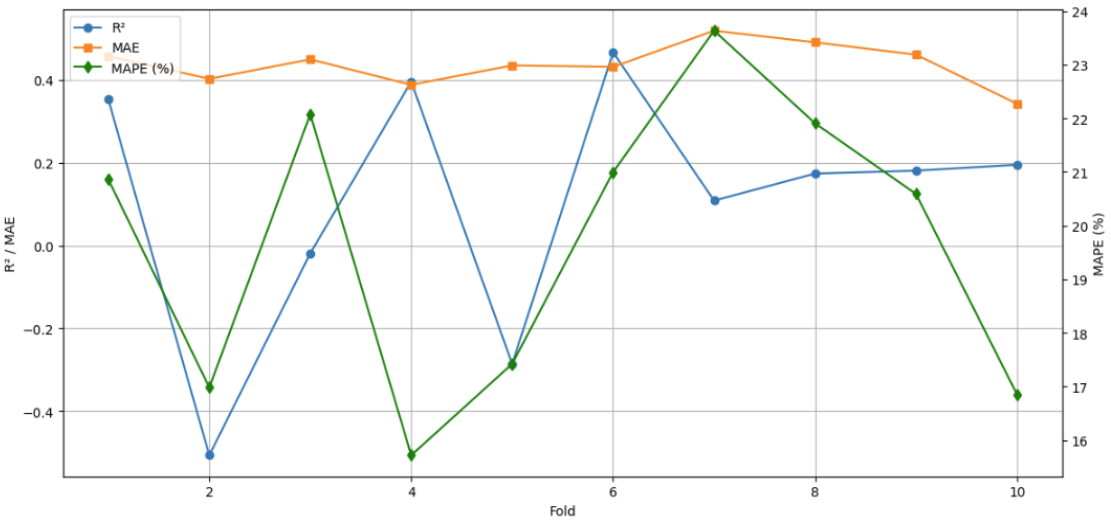


Fonte: O autor.

5.2.7 *Random Forest Regressor*

O algoritmo *Random Forest Regressor* apresentou desempenho insatisfatório na estimativa do COT. O R² médio de 0,1062, MAE de 0,4380 e MAPE de 19,71% indicam que apenas cerca de 10% da variabilidade da variável alvo (COT) foi explicada pelos modelos gerados. A Figura 25 apresenta o gráfico dos resultados obtidos com a validação cruzada.

Figura 25. Desempenho do algoritmo *Random Forest Regressor* para estimativa de COT

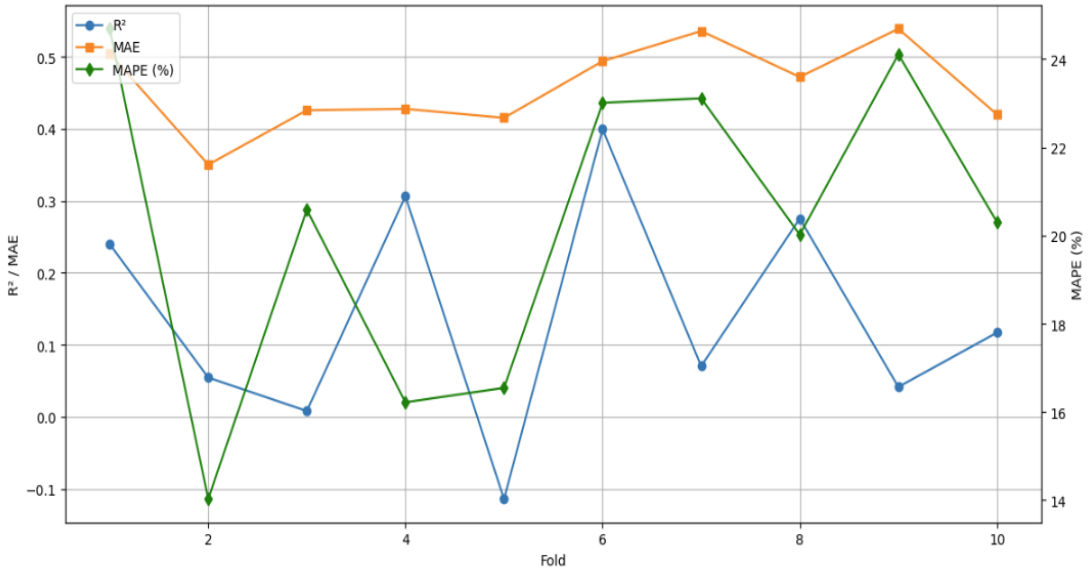


Fonte: O autor.

5.2.8 SVR

O algoritmo SVR também apresentou desempenho insatisfatório. Os valores médios de R^2 (0,1398), MAE (0,4582) e MAPE (20,26%) indicam que esse algoritmo não conseguiu criar um modelo capaz de predizer COT a partir das bandas espectrais contidas na base de dados (ver gráfico da Figura 26).

Figura 26. Valores de R^2, MAE e MAPE na validação cruzada do modelo gerado pelo algoritmo SVR para estimativa de COT

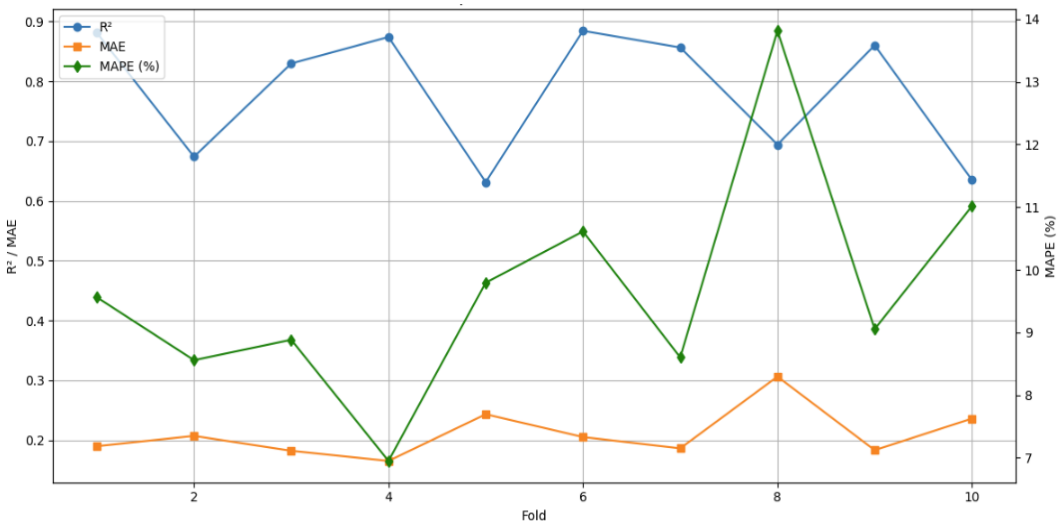


Fonte: O autor.

5.2.9 OMP

O algoritmo OMP apresentou desempenho promissor nos dados analisados. O R^2 foi de 0,78, indicando que 78% da variabilidade do atributo meta (COT) é explicada pelo modelo. O MAE foi de 0,21, indicando que, em média, as previsões diferem dos valores reais em cerca de 0,21 unidades, e o MAPE foi de 969%. Os detalhes do desempenho ao longo da validação cruzada podem ser vistos no gráfico da Figura 27.

Figura 27. Desempenho OMP para estimativa de COT



Fonte: O autor.

O Quadro 3 contém uma síntese dos resultados obtidos para a estimativa de COT, em ordem de melhor desempenho.

Quadro 3. Médias das métricas R^2, MAE e MAPE dos algoritmos de predição de COT.

Algoritmo	**R^2**	**MAE**	**MAPE (%)**
PLS Regression	0,8156	0,1994	9,24
Linear Regression + SPA	0,8003	0,2064	9,58
OMP	0,7823	0,2106	9,69
PLSR + SPA	0,7512	0,2246	10,28
SVR	0,1398	0,4582	20,26
Random Forest Regressor	0,1062	0,4380	19,71
Ridge Regression	0,0660	0,4895	22,00
MLP Regressor	0,0019	0,4996	22,22
Linear Regression	-1363,8743	11,4907	508,80

Fonte: O autor.

5.2.10 Desempenho na predição de COT

O modelo de melhor desempenho para a estimativa de COT também foi obtido por meio do algoritmo PLSR, que apresentou um R^2 médio de 0,8156. Esse resultado foi superior ao do trabalho de Dharumarajan et al. (2025), que usou PLSR tradicional e obteve um R^2 de 0,70. Com isso, pode-se considerar que o método de pré-processamento sequencial utilizado (transformação logarítmica, Z-score, filtro Savitzky–Golay) foi um fator-chave para a otimização da capacidade preditiva do PLSR. O modelo alcançou um desempenho estatisticamente similar ao do trabalho de Jakkan et al. (2024), que usou uma Rede Neural Artificial Sequencial (DrSeqANN) treinada com pré-processamento logarítmico, tendo como resultado R^2 de 0,82. No Quadro 4 tem-se uma síntese dessas comparações.

Quadro 4. Comparação do resultado do PLSR com trabalhos correlatos na estimativa de COT.

Modelo de Referência	**Estratégia Principal**	**R^2 de Predição (COT)**	**Comparação com resultado PLSR (R^2 = 0,8156)**
(Jakkan et al., 2024)	Deep Learning (DrSeqANN) + pré-processamento Log10x	0,82	O resultado de PLSR + pré-processamento sequencial é essencialmente igual ao de um modelo de *Deep Learning* mais complexo.
(Santos et al., 2020)	SVM ou PLSR em espectroscopia no infravermelho médio (MIR)	0,93	O resultado é inferior ao alcançado por Santos et al. (2020), o que indica maior precisão quando se utiliza a região MIR ou a combinação Vis-NIR + MIR.
(Dharumarajan et al., 2025)	PLSR em laboratório (Vis-NIR) sem agrupamento (estimativa de C)	0,70	O resultado é superior ao deste estudo, que utilizou um PLSR mais tradicional para estimar o teor de carbono.

Fonte: O autor.

6 CONCLUSÃO

O presente estudo propôs-se a avaliar a eficácia de diferentes abordagens de AM na estimativa dos teores de COT e de N em amostras de solo, com base em dados de espectroscopia NIR. A motivação central deste trabalho reside na busca por alternativas às análises químicas tradicionais, que, embora precisas, são frequentemente onerosas, demoradas e podem gerar resíduos.

Para atingir os objetivos propostos, utilizou-se uma base de dados com 235 amostras de solo, cada uma contendo os valores de referência de COT e de N, além de 277 atributos espectrais na faixa de 1100 a 1648 nm.

O pré-processamento dos dados constituiu uma etapa fundamental para garantir a qualidade e a robustez dos modelos. Foram aplicados os métodos de transformação logarítmica, de padronização Z-score, de normalização Min-Max e do filtro Savitzky-Golay. A avaliação dos modelos foi conduzida em duas frentes: uma utilizou a totalidade dos atributos espectrais e a outra empregou a técnica de seleção de variáveis SAP.

A robustez dos resultados foi assegurada por meio do uso da validação cruzada com 10 *folds*, e a performance final foi avaliada pelos valores médios de R^2, MAE e MAPE.

Na predição do teor de N, os resultados revelaram uma acentuada disparidade no desempenho dos algoritmos. Modelos lineares simples, como a *Linear Regression*, e modelos mais complexos, como o MLP *Regressor* e o *Random Forest Regressor*, apresentaram desempenho insatisfatório quando aplicados ao conjunto completo de variáveis, com valores de R^2 médios negativos ou próximos de zero (-9,1602, -0,0129 e 0,0896, respectivamente). Este resultado evidencia a incapacidade desses modelos de lidar com a alta dimensionalidade e a elevada similaridade dos valores nas faixas espectrais, um desafio inerente aos dados de espectroscopia. Em contrapartida, algoritmos especificamente projetados para tratar essas características, como o PLSR e o OMP, apresentaram desempenho notavelmente superior. O PLSR alcançou R^2 médio de 0,9124, MAE de 0,1253 e MAPE de 16,16%, sendo o melhor modelo, enquanto o OMP, com R^2 médio de 0,8327 e MAE de 0,1816, indicou uma importante capacidade de explicar a variabilidade do N a partir dos dados espectrais NIR.

A etapa de seleção de variáveis confirmou sua relevância. A aplicação do SPA antes do treinamento da Regressão Linear melhorou seu desempenho, resultando em um R^2 médio de 0,8975 nas 75 faixas espectrais selecionadas. Este é um dos achados mais significativos do trabalho, pois demonstra que a relação linear entre os espectros e o teor de nitrogênio existe, mas é ofuscada pela redundância e pelo ruído presentes nos dados brutos. A seleção de variáveis

não apenas simplificou o modelo, mas também o tornou o mais performático, superando até mesmo o OMP. Adicionalmente, a análise VIP no modelo PLSR identificou as faixas espectrais entre 1394 nm e 1412 nm como as mais relevantes para a predição de nitrogênio, fornecendo insights valiosos sobre quais regiões do espectro carregam a informação mais relevante.

De forma análoga, a predição do COT seguiu um padrão semelhante. Os algoritmos *Linear Regression* e o MLP *Regressor* falharam no conjunto completo de dados, obtendo valores de R^2 médios de 1363,8743 e 0,0019, respectivamente. O Random Forest e o SVR também apresentaram desempenho limitado. Os modelos PLSR e *Linear Regression* + SPA se sobressaíram, com R^2 médios de 0,8156 e 0,8003, respectivamente. Para a estimativa de COT, o PLSR também se consolidou como o algoritmo de melhor desempenho na base de dados completa (com 277 atributos de predição). A utilização da seleção de atributos com o SPA também impulsionou os resultados, com a Regressão Linear atingindo um R^2 de 0,8003 no subconjunto de 75 atributos. A métrica VIP no PLSR destacou a importância dos comprimentos de onda na faixa de 1396 nm a 1450 nm, reforçando que faixas espectrais específicas são cruciais para uma modelagem precisa.

A partir da análise comparativa e dos resultados obtidos, pode-se concluir que a combinação de espectroscopia NIR e AM é uma abordagem viável e de alto potencial para a quantificação de COT e de N no solo. O estudo demonstrou que a escolha do algoritmo é crucial; métodos robustos à multicolinearidade, como PLSR e OMP, são intrinsecamente mais adequados para dados espectrais de alta dimensionalidade. Por outro lado, o uso de seleção de atributos no pré-processamento transformou o desempenho dos algoritmos de Regressão Linear, mostrando que é possível usar um modelo linear simples em dados espectrais NIR, quando associados a uma quantidade menor e mais relevante de atributos.

Apesar dos resultados promissores, reconhecem-se limitações que abrem caminhos para pesquisas futuras. O estudo foi conduzido com um conjunto de dados de uma região específica, sendo necessária a validação dos modelos em solos de diferentes tipos e origens para avaliar sua capacidade de generalização. A exploração de algoritmos de *Deep Learning*, como *Redes Neurais Convolucionais* (CNNs), que podem aprender características espectrais de forma hierárquica, também representa uma fronteira promissora, especialmente com a disponibilidade de bases de dados maiores.

Este trabalho cumpre seu objetivo geral ao demonstrar que abordagens de aprendizado de máquina, combinadas a métodos adequados de pré-processamento, transformação de dados

e seleção de atributos, podem estimar satisfatoriamente os teores de COT e N a partir de dados espectrais NIR.

REFERÊNCIAS

AHMADI, A; EMAMI, M; DACCACHE, A; HE, L. Soil properties prediction for precision agriculture using visible and near-infrared spectroscopy: a systematic review and meta-analysis. **Agronomy**, v. 11, n. 433, p. 1-14, 2021.

AMORIM, L. B. V. de; CAVALCANTI, G. D. C.; CRUZ, R. M. O. The choice of scaling technique matters for classification performance. **Applied Soft Computing**, Amsterdam, v. 133, e109924, 2023.

CAGLIARINI, C.; ZINN, Y. L.; VALCARCEL, S.; COZZOLINO, D. Near Infrared Spectroscopy for On-Site Measurement of Total Organic Carbon in Soil: Effects of Moisture Content and Particle Size. **Sensors**, v. 17, n. 2366, 2017.

DHARUMARAJAN, S.; GOMEZ, C.; KUSUMA, C. G.; VASUNDHARA, R.; KALAISELVI, B.; LALITHA, M.; HEGDE, R. Prediction of soil organic carbon stock along layers and profiles using Vis-NIR laboratory spectroscopy. **Catena**, v. 257, p. 109150, 2025.

DONG, J.; WU, L. Comparison and Simulation Study of the Sparse Representation Matching Pursuit Algorithm and the Orthogonal Matching Pursuit Algorithm. **International Conference on Wireless Communications and Smart Grid** (ICWCSG), Hangzhou, China, 2021, pp. 317-320.

GAZELI, O.; STEFAS, D.; COURIS, S. Sulfur detection in soil by laser induced breakdown spectroscopy assisted by multivariate analysis. **Materials**, v. 14, n. 541, p. 1–16, 2021.

HE, S.; TAN, S.; SHEN, L.; ZHOU, Q. Soil Organic Matter Estimation Model Integrating Spectral and Profile Features. **Sensors** 2023, 23, 9868.

JAKKAN, D. A.; GHARE, P.; KUMAR, N.; SAKODE, C. Predicting the carbon property of soil using DrSeqANN and VIS/NIR spectroscopy. **Journal of Information Systems Engineering and Management**, v. 10, n. 8s, p. 1–12, 2025.

JAVAID, M; HALEEM, A; KHAN, I; SUMAN, R. Understanding the potential applications of Artificial Intelligence in Agriculture Sector. **Advanced Agrochem**, v. 2, p. 15–30, 2023.

JIN, X; ZHOU, J; RAO, Y; ZHANG, X; ZHANG, W; BA, W; ZHOU, X; ZHANG, T. An innovative approach for integrating two-dimensional conversion of Vis-NIR spectra with the Swin Transformer model to leverage deep learning for predicting soil properties. **Geoderma**, v. 436, p. 116555, 2023.

KUHN, M; JOHNSON, K. Applied Predictive Modeling. New York: **Springer**. 2013.

LARTEY, CLEMENT et al. Effective Outlier Detection for Ensuring Data Quality in Flotation Data Modelling Using Machine Learning (ML) Algorithms. **Minerals**, Basel, v. 14, n. 9, e925, 2024.

LIU, M.; YANG, K.; YAN, Y.; SONG, Z.; TIAN, F.; LI, F.; YU, Z.; ZHANG, R.; YANG, Q.; LU, Y. Multi-spectral evaluation of total nitrogen, phosphorus and potassium content in soil using Vis-NIR spectroscopy based on a modified support vector machine with whale optimization algorithm. **Soil & Tillage Research**, v. 252, p. 106567, 2025.

LUO, R; POPP, J; BOCKLITZ, T. Deep Learning for Raman Spectroscopy: A Review. **Analytica**, v. 3, p. 287–301, 2022.

MARSCHNER, H. Marschner's Mineral Nutrition of Higher Plants (3rd ed.). **Academic Press**. 2012.

MATAMALA, R.; JASTROW, J. D.; CALDERÓN, F. J.; LIANG, C.; FAN, Z.; MICHAELSON, G. J.; PING, C. L. Predicting the decomposability of arctic tundra soil organic matter with mid infrared spectroscopy. **Soil Biology and Biochemistry**, v. 129, p. 1–12, 2019.

MIRANI, F.; MAFFINI, A.; CASAMICHIELA, F.; PAZZAGLIA, A.; FORMENTI, A.; DELLASEGA, D.; RUSSO, V.; VAVASSORI, D.; BORTOT, D.; HUAULT, M.; ZERAOULI, G.; OSPINA, V.; MALKO, S.; APIÑANIZ, J. I.; PÉREZ-HERNÁNDEZ, J. A.; DE LUIS, D.; GATTI, G.; VOLPE, L.; POLA, A.; PASSONI, M. Integrated quantitative PIXE analysis and

EDX spectroscopy using a laser-driven particle source. **Science Advances**, v. 7, n. 3, eabc8660, 2021.

MUNAWAR, A. A.; YUNUS, Y.; DEVIANTI; SATRIYO, P. Calibration models database of near infrared spectroscopy to predict agricultural soil fertility properties. **Data in Brief**, v. 30, p. 105469, 2020.

NAWAR, S.; MOUAZEN, A. M. Predictive performance of mobile vis–near infrared spectroscopy for key soil properties at different geographical scales by using spiking and data mining techniques. **Catena**, v. 151, p. 118–129, 2017.

PASQUINI, C. Near infrared spectroscopy: a mature analytical technique with new perspectives – a review. **Analytica Chimica Acta**, v. 1026, p. 8-36, 2018.

PUDEŁKO, A; CHODAK, M; ROEMER, J; UHL, T. Application of FT-NIR spectroscopy and NIR hyperspectral imaging to predict nitrogen and organic carbon contents in mine soils. **Measurement**, v. 164, p. 108117, 2020.

RAJAN, M.P. An Efficient Ridge Regression Algorithm with Parameter Estimation for Data Analysis in Machine Learning. SN **COMPUT**. SCI. 3, 171 (2022). https://doi-org.ez82.periodicos.capes.gov.br/10.1007/s42979-022-01051-x.

RAYMAEKERS, Jakob; ROUSSEEUW, Peter J. Transforming variables to central normality. **Machine Learning**, v. 113, p. 4953-4975, 2024.

RINNAN, Å.; VAN DER BERG, F. W. J.; ENGELSEN, S. B. (2009). Review of the most common pre-processing techniques for near-infrared spectra. TrAC, Trends in Analytical **Chemistry**, 28(10), 1201-1222.

SANTOS, U. J. dos; DEMATTÊ, J. A. de M.; MENEZES, R. S. C.; DOTTO, A. C.; GUIMARÃES, C. C. B.; ALVES, B. J. R.; PRIMO, D. C.; SAMPAIO, E. V. de S. B. Predicting carbon and nitrogen by visible near-infrared (Vis-NIR) and mid-infrared (MIR) spectroscopy in soils of Northeast Brazil. **Geoderma Regional**, v. 23, e00333, 2020.

SCHMID, Michael; RATH, David; DIEBOLD, Ulrike. Why and How Savitzky-Golay Filters Should Be Replaced. **ACS Measurement Science Au**, Washington, DC, v. 2, p. 185-196, 2022.

SCHMIDT, M P.; SICILIANO, S D.; PEAK, D. The role of monodentate tetrahedral borate complexes in boric acid binding to a soil organic matter analogue. **Chemosphere**, v. 276, p. 130150, 2021.

TAMBURINI, E; VINCENZI, F; COSTA, S; MANTOVI, P; PEDRINI, P; CASTALDELLI, G. Effects of moisture and particle size on quantitative determination of total organic carbon (TOC) in soils using near-infrared spectroscopy. **Sensors**, v. 17, n. 2366, p. 1-15, 2017.

TAO, Z.; TAO, A.; LU, Y.; LI, X.; LIU, F.; KONG, W. Integrating fusion strategies and calibration transfer models to detect total nitrogen of soil using Vis-NIR spectroscopy. **Chemosensors**, v. 13, p. 1–19, 2025.

TORSONI, G; APARECIDO, L; SANTOS, G; CHIQUITTO, A; MORAES, J; ROLIM, G. Soybean yield prediction by machine learning and climate. **Theoretical and Applied Climatology**, v. 151, p. 1709–1725, 2023.

VAN EYND, E; WENG, L; COMANS, R N.J. Boron speciation and extractability in temperate and tropical soils: A multi-surface modeling approach. **Applied Geochemistry**, v. 123, p. 104797, 2020.

YE, N. The Handbook of Data Mining. **Lawrence Erlbaum Associates, Incorporated**, 2004.

ZHANG, X.; XUE, J.; XIAO, Y.; SHI, Z.; CHEN, S. Towards. Optimal Variable Selection Methods for Soil Property Prediction Using a Regional Soil Vis-NIR Spectral Library. **Remote Sens**. 2023, 15, 465. https://doi.org/10.3390/rs15020465.

ZHAO, W; LU, B; YU, J; ZHANG, B; ZHANG, Yuan. Determination of sulfur in soils and stream sediments by wavelength dispersive X-ray fluorescence spectrometry. **Microchemical Journal**, v. 156, p. 104840, 2020.

ZHENG, L.; WANG, Y.; LI, M.; LIM, Z.; JESSIC, R. J. Comparison of soil total nitrogen content prediction models based on Vis-NIR spectroscopy. **Sensors**, v. 20, n. 7078, p. 1-20, 2020.

ZHU, J; JIN, X; LI, S; HAN, Y; ZHENG, W. Prediction of soil available boron content in visible-near-infrared hyperspectral based on different preprocessing transformations and characteristic wavelengths modeling. **Computational Intelligence and Neuroscience**, v. 2022, p. 1-16, 2022.