UNIVERSIDADE ESTADUAL DE PONTA GROSSA
SETOR DE ENGENHARIAS, CIÊNCIAS AGRÁRIAS E TECNOLOGIA
PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO APLICADA

GABRIEL PASSOS DE JESUS

APRENDIZADO DE MÁQUINA EM DADOS DE ESPECTROSCOPIA DO INFRAVERMELHO PRÓXIMO PARA ESTIMATIVA DO TEOR DE CARBONO ORGÂNICO TOTAL NO SOLO

PONTA GROSSA
2025

GABRIEL PASSOS DE JESUS

APRENDIZADO DE MÁQUINA EM DADOS DE ESPECTROSCOPIA DO INFRAVERMELHO PRÓXIMO PARA ESTIMATIVA DO TEOR DE CARBONO ORGÂNICO TOTAL NO SOLO

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

Orientador: Prof. Dr. Eduardo Fávero Caires
Coorientador(a): Profa. Dra. Alaine Margarete Guimarães

PONTA GROSSA

2025

Jesus, Gabriel Passos de
J58 Aprendizado de máquina em dados de espectroscopia do infravermelho próximo para estimativa do teor de carbono orgânico total no solo / Gabriel Passos de Jesus. Ponta Grossa, 2025.
71 f.

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

Orientador: Prof. Dr. Eduardo Fávero Caires.
Coorientadora: Profa. Dra. Alaine Margarete Guimarães.

1. Espectroscopia. 2. Solos tropicais. 3. Aprendizado de máquina. 4. Carbono orgânico total. 5. Modelos regressivos. I. Caires, Eduardo Fávero. II. Guimarães, Alaine Margarete. III. Universidade Estadual de Ponta Grossa. Computação para Tecnologias em Agricultura. IV.T.

CDD: 004

Ficha catalográfica elaborada por Maria Luzia Fernandes Bertholino dos Santos- CRB9/986

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

Gabriel Passos de Jesus

"APRENDIZADO DE MÁQUINA EM DADOS DE ESPECTROSCOPIA DO INFRAVERMELHO PRÓXIMO PARA ESTIMATIVA DO TEOR DE CARBONO ORGÂNICO TOTAL 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:

Prof. Dr. Eduardo Fávero Caires (UEPG/PR - Presidente)

Prof. Dr. Arion de Campos Júnior (UEPG/PR)

Prof. Dr. André Luiz Lopes de Faria (DGE/UFV-MG)

Ponta Grossa, 25 de fevereiro de 2025.

Documento assinado eletronicamente por **Arion de Campos Junior, Professor(a)**, em 25/02/2025, às 10:16, conforme Resolução UEPG CA 114/2018 e art. 1º, III, "b", da Lei 11.419/2006.

Documento assinado eletronicamente por **Eduardo Favero Caires, Professor(a)**, em 25/02/2025, às 10:23, conforme Resolução UEPG CA 114/2018 e art. 1º, III, "b", da Lei 11.419/2006.

Documento assinado eletronicamente por **André Luiz Lopes de Faria, Usuário Externo**, em 26/02/2025, às 14:48, 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 **2421577** e o código CRC **B5607E7C**.

Dedico esta dissertação à minha noiva Helena Pistune que me incentivou a não desistir da carreira acadêmica e tanto me apoiou nos momentos difíceis. Te amo!

AGRADECIMENTOS

À CAPES – Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – pelo suporte financeiro.

À UEPG, Universidade Estadual de Ponta Grossa, e ao Programa de Pós-Graduação em Computação Aplicada, pela oportunidade de cursar o Mestrado.

À Fundação ABC e MSc. Grazielle Schornobai pelo fornecimento das amostras de solo para a realização deste estudo.

À minha preciosa Helena Pistune que tanto me incentivou a cursar um mestrado, além de agradecer por todos os momentos desde que nos conhecemos. Obrigado por tudo, te amo!

Aos meus pais Rubens e Maria D'Ajuda, ao meu irmão Rafael, por todo o suporte desde que saí de casa para estudar fora aos 16 anos.

À minha sogra Joana D'Arc que se tornou uma segunda mãe; à minha cunhada Elisa Maria que se tornou minha irmã mais nova; à minha tia Ana Paula pelo acolhimento.

Ao meu amigo Cleverson Costa pelo acolhimento em terras paranaenses, pelas conversas sobre futebol e automobilismo e por sanar minhas dúvidas em Química.

Aos meus familiares e amigos que, mesmo de longe, torcem pelo meu sucesso.

Aos meus orientadores, Prof. Dr. Eduardo Fávero Caires e Profa. Dra. Alaine Margarete Guimarães pelos ensinamentos. Estendo o agradecimento aos colegas do LabFer – Laboratório de Fertilidade do Solo – e InfoAgro – Laboratório de Informática Aplicada às Ciências Agrárias e Ambientais.

Aos professores do Programa de Pós-Graduação em Computação Aplicada pelo acolhimento, paciência e suporte. Agradeço nominalmente aos professores Drª. Maria Salete Marcon Gomes Vaz, Dr. Arion de Campos Jr., Dr. Alceu de Souza Britto Jr., Dr. Adriano Ferrasa, Dr. José Carlos Ferreira da Rocha, Dr. Luciano José Senger, Dr. Zito Palhano da Fonseca, Dra. Rosane Falate, Dr. Sérgio Luiz Stevan Jr, Dr. Jonathan de Matos e Dr. Rafael Mazer Etto, por todo o carinho e cuidado que tiveram ao me receber no programa.

Aos amigos que fiz pelo caminho, em especial Bruno Seixas Bonfati, Emili Everz Golombieski, Danilo Alves da Rosa, Maria Beatriz Macedo Simon Sola, Mateus "Gilson" Felipe Ribeiropelo apoio, pelas conversas, dicas e suporte na execução deste trabalho.

Aos meus amigos Gabriel Tonon Cimatti e Caroline Heloíse de Oliveira que tive a oportunidade de coorientá-los em seus trabalhos de Iniciação Científica.

Aos meus amigos da UFV: Filipe Abdala Bento e Dantas de Moraes pelas horas e horas conversando sobre Engenharia e Tecnologia. Pelas diversas listas de Cálculo e Álgebra Linear. Pelos memes compartilhados e por toda a parceria desde 2018; Rafaella Barroso Vieira Rabello pela parceria desde as aulas de DIR 130 e Rayssa Oliveira pela parceria desde as monitorias de Fundamentos da Geografia. A vocês, muito obrigado pelo carinho e apoio na execução deste trabalho.

À minha amiga Cris Ikki Dracco pelos memes, fotos de animais e conversas sobre Física e futebol. aos meus padrinhos de YA, Danielly Felippi, Miguel Fabrin e Danilo Ricardo por todo o apoio desde 2015.

A minha querida Lilian Cordeiro por todo o suporte.

Aos meus amigos Ricardo Razera, Luiz Miranda, Luiz Salmeron, Marcelo Modesto, Juliana Betti, Nicolas Leme – e a todos do Projeta Cidadania, de Sorocaba, SP.

Aos meus amigos do Programa de Pós-Graduação em Agronomia, em especial, Luiz Carlos Costa, Calistene Aparecida Pinto, Ranyelly Leão, Waldir Zarrochinski Jr., Jean-Fritz Milien e Thiago Guse, pelo apoio na realização dos experimentos, pelas conversas agradáveis e diversos momentos de companheirismo e risadas!

Aos demais amigos que fiz em diversos cursos da Universidade Estadual de Ponta Grossa.

Aos meus amigos da graduação em Geografia da Universidade Federal de Viçosa, em especial ao Marco Túlio Cunha, Renato Pereira, Thulio Caprunio, Carla Martins, Carina Teixeira, Marco Antônio Saraiva da Silva, Yann Xaneis e aos professores que me formaram na graduação, em especial, aos meus grandes amigos Prof. Dr. Leonardo Civale, Dr. André Luiz Lopes de Faria e Dr. Edson Soares Fialho, por serem meus pais científicos.

Ao meu professor e amigo Dr. Carlos Tadeu dos Santos Dias (ESALQ/USP) pelas conversas e dicas sobre Estatística.

Aos professores Dr. Arion de Campos Jr, Drª. Karen Wohnrath, Dr. José Carlos Ferreira da Rocha e Dr. André Luiz Lopes de Faria, que, gentilmente, aceitaram compor a banca examinadora desta dissertação. Suas contribuições são muito válidas.

Ao povo brasileiro que contribuiu direta e indiretamente para a realização deste Mestrado.

“Just like moons and like suns, With the certainty of tides, Just like hopes springing high, Still I'll rise.”

“Assim como as luas e os sóis, Com a certeza das marés, Assim como as esperanças brotam altas, Ainda assim eu vou me levantar.”

- Maya Angelou

RESUMO

Com o advento da agricultura de precisão, novas tecnologias passaram a ser utilizadas no campo. Nos últimos anos, houve expansão de correntes teóricas e novas práticas que contribuem para uma agricultura mais sustentável e limpa, por meio de soluções para processos que, até então, eram considerados poluentes para o meio ambiente e onerosos financeiramente. Assim, a utilização de técnicas de espectroscopia – com ênfase no NIRS – aliada às técnicas de aprendizado de máquina têm sido vistas como alternativa para dirimir os efeitos causados pela poluição com reagentes químicos, bem como a diminuição de custos operacionais. O presente trabalho tem como objetivo geral a determinação de modelos de estimativa dos teores de carbono orgânico no solo com a utilização de técnicas de aprendizado de máquina. O trabalho tem como objetivos específicos a construção de bases de dados de leitura espectral na região do infravermelho, a identificação de técnicas de aprendizado de máquina que se mostrem adequadas para o perfil das bases de dados criadas e avaliar, por meio do coeficiente R^2, o potencial do uso dos modelos gerados. As amostras de solo para este estudo foram coletadas de um experimento de longa duração (pouco mais de 30 anos) realizado em Ponta Grossa, Paraná. Para este trabalho foram utilizados seis algoritmos, a saber: Huber Regressor, Extra Trees Regressor, Catboost Regressor, ElasticNet, Lasso e Bayesian Ridge Regressor. O algoritmo Huber Regressor se destacou como melhor, considerando a base quadrática ($R^2 = 0,86$) e base padrão sem transformação ($R^2 = 0,66$). Já, para a base logarítmica, o maior destaque foi do algoritmo Extra Trees Regressor ($R^2 = 0,54$).

Palavras-chave: Espectroscopia, Solos Tropicais, Aprendizado de Máquina, Carbono Orgânico Total, Modelos Regressivos

ABSTRACT

With the advent of precision agriculture, new technologies have been adopted in the field. In recent years, theoretical frameworks and innovative practices have expanded, contributing to more sustainable and eco-friendly agriculture by offering solutions for processes previously deemed environmentally harmful and financially burdensome. Thus, the use of spectroscopy techniques—particularly NIRS—combined with machine learning methods has emerged as an alternative to mitigate pollution and reduce operational costs. This study aimed to develop predictive models for estimating soil organic carbon content using machine learning techniques. Specific objectives include building spectral data sets in the infrared region, identifying suitable machine learning techniques for the created datasets, and evaluating the potential of the generated models using the R^2 coefficient. Soil samples for this study were taken from a long-term experiment lasting just over 30 years in Ponta Grossa, Parana State, Brazil. Six algorithms were employed in this study: Huber Regressor, Extra Trees Regressor, Catboost Regressor, ElasticNet, Lasso, and Bayesian Ridge Regressor. The Huber Regressor algorithm stood out as the best, considering the quadratic basis ($R^2 = 0.86$) and standard basis without transformation ($R^2 = 0.66$). For the logarithmic basis, the greatest highlight was the Extra Trees Regressor algorithm ($R^2 = 0.54$).

Keywords: Spectroscopy, Tropical Soils, Machine Learning, Total Organic Carbon, Regression Models

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE TABELAS

LISTA DE SIGLAS

AM	Aprendizado de Máquina
BRR	Regressão Bayesian Ridge
COT	Carbono Orgânico Total
EMR	Radiação Eletromagnética
IA	Inteligência Artificial
KDD	Descoberta do Conhecimento por Bases de Dados
LASSO	Operador de seleção e encolhimento mínimo absoluto
MAE	Erro Absoluto Médio
NIR	Infravermelho Próximo
RMSE	Raiz do Erro Quadrático Médio

SUMÁRIO

1 INTRODUÇÃO

A produção de alimentos requer solos férteis, contudo, práticas intensivas podem degradá-los. Dessa forma, a preocupação com a preservação do ecossistema do solo ganhou espaço nos estudos científicos (Angelopoulou *et al*., 2020). Em anos recentes, o carbono tornou-se um dos temas centrais no estudo da ciência do solo (Barra *et al.*, 2021) por ser indicador-chave de qualidade do solo. A presença de carbono traz consequências que impactam diretamente a fertilidade e a saúde do solo (Gerke, 2022). De acordo com Seema (2022), a concentração de carbono orgânico total (COT) é um indicador ligado às funções essenciais do solo. Sua importância relaciona-se com a ciclagem, armazenamento e absorção de nutrientes, bem como na retenção de poluentes.

Métodos convencionais de análise, como a quimiometria por dicromato de potássio ($K_2Cr_2O_7$) – também chamado de "análise por via úmida" – são altamente poluentes, tanto para sistemas de esgoto como para o solo e a atmosfera (Mostert *et al.*, 2010). Desta maneira, a utilização de técnicas alternativas, como a espectroscopia, para a estimativa dos teores de COT, bem como suas propriedades, tem se tornado amplamente difundida no estudo da ciência do solo e da agricultura de precisão.

A utilização da espectroscopia corresponde ao uso de equipamentos que atuam em diferentes comprimentos de onda, como o infravermelho (IR – entre 780 e 1.000.000 nm) e dentro desse, o infravermelho próximo (NIR – cerca de 780 a 2.500 nm). A radiação infravermelha é uma radiação eletromagnética (EMR) com comprimentos de onda mais longos do que os da luz visível. É, portanto, invisível para o olho humano, mas pode ser percebida na forma de radiação térmica. Quando a radiação infravermelha é direcionada ao alvo (amostra que está sendo analisada), pode estimular o movimento de moléculas e ligações atômicas. Dependendo de como a molécula está excitada pode-se obter informações sobre as características do material irradiado. Substâncias orgânicas e inorgânicas podem ser examinadas igualmente, bem com radiação infravermelha, sendo que o requisito básico é que o material absorva tal radiação, como é o caso do solo (Barra *et al*, 2021).

O uso da espectroscopia apresenta vantagens em relação aos métodos convencionais por diminuir a quantidade de poluentes, bem como reduzir o tempo de realização do procedimento e os custos operacionais.

O uso da espectroscopia implica na necessidade da criação de modelos de estimativa do elemento-alvo. Esses modelos são obtidos e calibrados, via de regra, com base em métodos estatísticos e computacionais aplicados a conjunto de dados que correlacionam a resposta

espectral com valores de referência obtidos por meio de análises convencionais em laboratório (Ramírez *et al.*, 2020). Outra forma de se obter tais modelos é com o uso de técnicas de Aprendizado de Máquina (AM). O AM é um ramo da Inteligência Artificial (IA) que se concentra no desenvolvimento de algoritmos e modelos que permitam aos sistemas computacionais aprender a partir de dados e melhorarem seu desempenho nas tarefas sem serem explicitamente programados (Russel; Norvig, 2003).

Na agricultura de precisão, o uso de IA desempenha um papel fundamental na gestão do solo em todas as suas etapas de manejo (semeadura, cultivo, colheita e pós-colheita) (Eli-Chukwu, 2019; Karunathilake *et al.*, 2023). Mais especificamente, o AM já vem sendo utilizado na agricultura no reconhecimento de padrões de manejo de doenças em plantas, invasão de plantas-daninhas, estimativa de produtividade de cultivares e na mensuração dos teores de compostos presentes no solo (Mucherino; Papapjorgi; Pardalos, 2009).

2 OBJETIVOS

2.1 Objetivo Geral

O presente trabalho tem como objetivo geral a determinação de modelos de estimativa dos teores de COT no solo por meio de espectroscopia na região do NIR com a utilização de técnicas de AM.

2.2 Objetivos Específicos

- Construir uma base de dados de leitura espectral na região do infravermelho de amostras do solo integrada com as respectivas análises em laboratório do teor de COT.
- Identificar técnicas de Aprendizado de Máquina que se mostrem adequadas para o perfil da base de dados criada.
- Avaliar, por meio de métricas estatísticas, com ênfase no coeficiente de determinação (R^2), o potencial de uso dos modelos computacionais gerados.

3 REVISÃO DA LITERATURA

3.1 Agricultura e Espectroscopia

Os estudos das temáticas agrícolas têm ganhado espaço, sobretudo as discussões envolvendo a presença de carbono no solo (Barra *et al*., 2021; Yang *et al*., 2022). Há também a preocupação quanto a emissão de carbono na atmosfera (Natali *et al*, 2019; Santana *et al*, 2019; Zhou *et al*, 2023), bem como a preocupação de se ter melhor tomada de decisão no campo (Angelopoulou *et al*, 2020; Navarro; Costa; Pereira, 2020). Acerca do carbono, segundo Gerke (2022), a sua emissão na atmosfera possui consequências que impactam diretamente na saúde do solo, afetando processos de ciclagem, erosão e degradação, reduzindo a fertilidade do solo. A concentração do COT, segundo Seema (2020), é um indicador-chave da qualidade ligada a funções essenciais do ecossistema do solo, incluindo ciclagem e armazenamento de nutrientes.

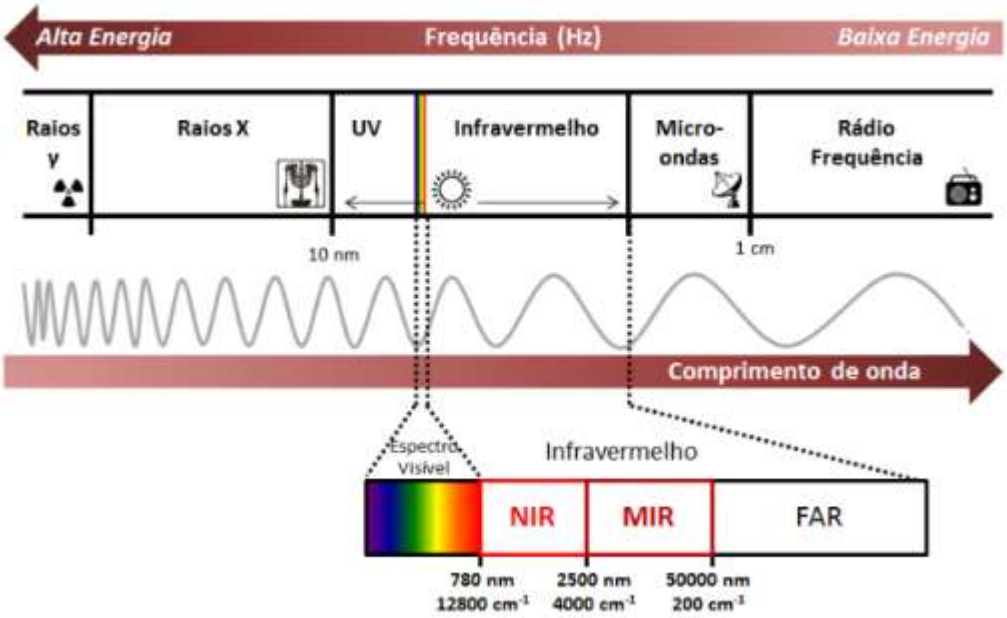
Métodos convencionais de análise, como a quimiometria por dicromato de potássio (K_2Cr_2O7) – também chamado de "análise por via úmida" – são altamente poluentes tanto para sistemas de esgoto, como para o solo e a atmosfera (Mostert *et al*, 2010). Destarte, a utilização de técnicas alternativas, como a espectroscopia, para a estimativa dos teores de COT, bem como suas propriedades, tem se tornado amplamente difundida no estudo da ciência do solo e da agricultura de precisão.

3.2 Utilização da Espectroscopia

A espectroscopia é uma técnica que, em relação a outros procedimentos, apresenta as seguintes vantagens: (i) simplicidade na preparação das amostras de solo, dispensa da necessidade de armazenagem em condições especiais, (ii) possibilidade de se determinar diversos parâmetros do solo em pouco tempo, (iii) menor custo de análise e (iv) redução (ou não geração) de resíduos poluentes (Ahmadi *et al*., 2021; Hati *et al*., 2022). Desta maneira, tem sido utilizada como uma técnica de análise não-destrutiva (Marchão *et al*., 2011; Vashisth *et al*., 2022).

A espectroscopia no infravermelho se baseia na ideia de que ligações químicas das substâncias possuem frequências de vibração específicas, as quais correspondem a níveis de energia da molécula (Whetsel, 1968; Jung; Jung; Cole, 2023). A região do infravermelho (IR) estende-se dos $3x10^{11}$ Hz até aproximadamente os $4x10^{14}$ Hz e é subdividida em três regiões: o NIR, também chamado IR-próximo (i.e., próximo da luz visível, ~780 – 2.500 nm), o IR-médio (MIR, ~2.500 – 50.000 nm) e o IR-longoFAR, ~50.000 nm – 1.000.000 nm) (Figura 01).

Figura 01. Espectro Eletromagnético



Fonte: Eskildsen (2016)

A espectroscopia no infravermelho próximo (NIRS) é uma técnica analítica utilizada para identificar e quantificar compostos químicos presentes em amostras (Ramírez *et al.*, 2021). É baseada na interação entre a radiação eletromagnética na faixa do infravermelho próximo e as vibrações moleculares das substâncias (Gholizadeh *et al.*, 2021). Ao incidir na amostra de solo a radiação na faixa do infravermelho próximo, ocorrem processos de penetração, refração e reflexão múltiplas, gerando radiação difusa. Essa radiação capturada fornece informações sobre a composição da amostra sendo analisada (Barbosa, 2020).

O esquemático (Figura 02) exibe o processo básico da espectroscopia:

Figura 02. Esquemático da Espectroscopia

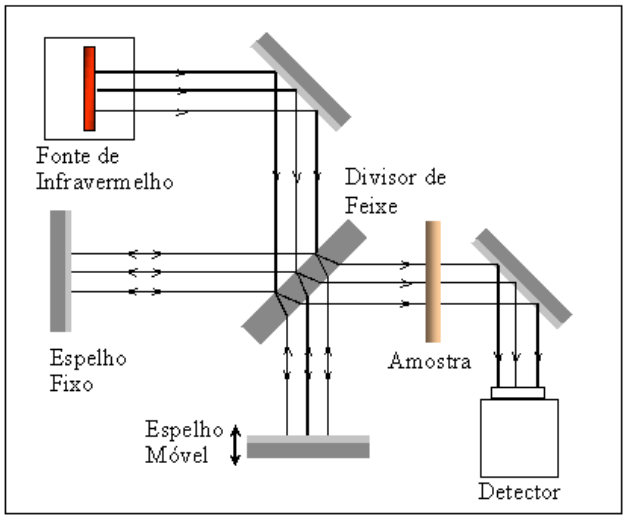


Imagem: Unesp (reprodução)

Cada material (amostra) submetido ao espectrofotômetro apresenta uma assinatura espectral. O espectro obtido pode ser analisado de várias maneiras, dependendo do objetivo da análise. A partir do espectro, é possível identificar diferentes espécies químicas, estudar interações moleculares, determinar a estrutura molecular, investigar propriedades físicas, como a absorção e emissão de luz, e quantificar a concentração de compostos, sendo esta, a característica de principal interesse para esse trabalho.

3.3 Espectroscopia e Aprendizado de Máquina

A integração da espectroscopia com algoritmos de AM tem potencializado a capacidade de análise e interpretação dos dados de solos (Wadoux, 2023). Modelos preditivos, tais como os modelos regressivos, por exemplo (Ribeiro *et al*., 2021), treinados com grandes conjuntos de dados espectrais e de referência de solos, permitem a predição precisa das propriedades do solo a partir de suas assinaturas espectrais (Padarian; Minasny; McBratney, 2019).

Isso não só acelera o processo de análise, mas também reduz os custos associados aos métodos laboratoriais tradicionais. Essas inovações têm contribuído para práticas agrícolas mais sustentáveis, permitindo a aplicação otimizada de insumos como fertilizantes e água, com base nas necessidades específicas de diferentes áreas do campo (Jain; Ramesh; Battacharya, 2021).

3.4 Mineração de Dados e Processos de KDD

A mineração de dados, do inglês *data mining*, é um campo interdisciplinar que se concentra na descoberta de padrões, tendências e relacionamentos em grandes conjuntos de dados (Salloum *et al*., 2021). Utilizando técnicas de AM, estatística e banco de dados (Li; Zhou; Cao, 2021), a mineração de dados transforma dados brutos em informações significativas que podem ser usadas para tomada de decisões em diversas áreas, como agricultura, negócios, saúde, finanças e ciências sociais (Tan *et al*., 2009). Esse processo, conhecido como Descoberta de Conhecimento em Bases de Dados (*Knowledge Discovery in Databases* ou chamado de KDD), envolve várias etapas, incluindo coleta de dados, limpeza, integração, seleção de atributos, mineração propriamente dita e avaliação dos resultados obtidos (Figura 03).

Figura 03 - Processo de mineração de dados em KDD

Fonte: Adaptado de Fayad et al (1996)

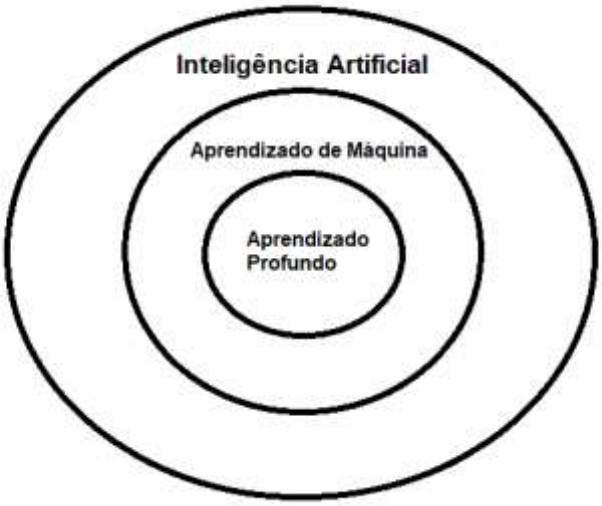
O processo de KDD, conforme descrito por Fayyad (1996) e Plotnikova *et al* (2020), consiste em 5 etapas. A primeira etapa é a seleção dos dados a partir da obtenção dos dados brutos. Após este momento, os dados passam pela etapa de pré-processamento, que é quando os dados-alvo começam a ser manipulados. A terceira etapa compreende o processo de transformação, normalmente por meio de métodos estatístico-matemáticos. A quarta etapa é a mineração propriamente dita, onde os padrões de relacionamento são estabelecidos para que ocorra a última etapa, que é a obtenção das bases de conhecimento; em outras palavras, *insights.*

Um desafio significativo na mineração de dados é a "maldição da dimensionalidade" (Bellman, 1957, 1961; Márquez, 2022). Esse termo refere-se aos problemas que surgem quando se trabalha com dados de alta dimensionalidade, ou seja, conjuntos de dados com um grande número de atributos (variáveis). À medida que o número de dimensões aumenta, o volume do

espaço de dados cresce exponencialmente, o que pode levar a dificuldades em encontrar padrões significativos (Chandra *et al*., 2023). Problemas comuns incluem o aumento da esparsidade dos dados, onde a correlações entre os atributos se tornam mais difíceis de serem identificadas e a degradação do desempenho dos algoritmos, que podem se tornar ineficazes ou computacionalmente inviáveis, tornando, segundo Camargo (2010), o processo de análise dos dados ainda mais complexo.

O AM (Figura 04) que se concentra no desenvolvimento de algoritmos e técnicas que permitem a predição e a tomada de decisões com base em dados (van Klopenburg; Kassahun; Catal, 2020). Em vez de seguir instruções programadas explicitamente, os algoritmos de AM identificam padrões e preveem resultados futuros ou classificam informações novas. O AM pode ser dividido em quatro tipos, sendo: supervisionado, não-supervisionado, semi-supervisionado e por reforço (Janiesch; Zschech; Heinrich, 2021).

Figura 04 – Diagrama das áreas da Inteligência Artificial



Fonte: Elaborado pelo próprio autor

No aprendizado supervisionado, os algoritmos aprendem a partir de dados rotulados, ou seja, tanto os dados de entrada como os de saída (rótulos) são previamente conhecidos *(input* e *output)* (Li; Guo; Zhou, 2021). Alguns dos modelos são baseados em árvores, ao passo que outros são baseados em regressões (Huang *et al*., 2023). O aprendizado não-supervisionado identifica padrões de agrupamento *(clustering)*. Já o aprendizado semi-supervisionado é uma combinação de dados rotulados e não rotulados (Stepisnik; Kocev, 2021). Por fim, o aprendizado por reforço visa a tomada de decisão por recompensa ou penalidade (mais utilizado na robótica ou em jogos) (Sharma; Srinvias; Ravindran, 2023).

3.5 Aprendizado de Máquina e Agricultura

Na agricultura, as aplicações do AM passaram a ganhar força entre os anos 1990 e 2000 (Liakos *et al.*, 2018) ao utilizar suas técnicas para a predição de parâmetros dos atributos do solo, análise da previsão de produtividade, manejo de recursos, prevenção de doenças, entre outras aplicações (Araújo *et al*., 2023).

Neste trabalho, cujo objetivo foi estimar os teores de COT com base em dados de espectroscopia NIR, o que implica na geração de base de dados de alta dimensionalidade, foram escolhidos seis algoritmos para a análise, sendo eles, baseados na tarefa de regressão, que apresentam bom desempenho no tratamento de bases de dados de alta dimensionalidade e com o objetivo de evitar o *overfitting* (sobreajuste do modelo), como o caso do Regressor de Huber (Huber, 1964; Feng; Wu., 2022) e do Catbooster Regression (Dorogush; Ershov; Gulin, 2018). O primeiro algoritmo citado tem como objetivo a garantia da sensibilidade dos dados (Sun; Zhou; Fan, 2019) de forma que o segundo é conhecido por sua capacidade de generalização (Zhang; Wang; Zhao, 2021). Além disso, foi utilizado o algoritmo Extra Trees Regressor. O algoritmo proposto por Geurts, Ernst e Wehenkel (2006) é uma derivação do Random Forest (Breiman, 2001). O Extra Trees é um regressor projetado para aumentar a eficiência computacional e diminuir a complexidade da operação, sendo um *ensemble* (conjunto) de árvores de decisão (Carrizosa; Molero-Río; Morales, 2021). Foram utilizados ainda os seguintes algoritmos: Elastic Net, (*Least Absolute Shrinkage and Selection Operator*) (Tibshirani, 1996) e Regressão *Bayesian Ridge* (MacKay, 1992). A escolha destes três algoritmos se deu pelo fato do Elastic Net ser um algoritmo que combina os outros citados por meio de penalização mista. Ou seja, de uma adoção equilibrada dos pesos das variáveis. Desta maneira, o Elastic Net se mostra como uma opção para garantir que a forte correlação entre as variáveis não ocorra apenas devido a um processo de overfitting. Assim, o algoritmo é uma contraposição ao fato do LASSO selecionar uma única variável e ignorar as demais (Frejeiro-González; Winch; González-Manteiga, 2021) e de uma possível distribuição de pesos de forma ineficiente, como pode ocorrer no Ridge Regression (Jenzen; Ramírez, 2008). Vale lembrar que, apesar do nome, o algoritmo Elastic Net não é oriundo da ideia de redes neurais, embora apresente um processo matemático similar (Zhang; Dai; Wu, 2022). Os algoritmos serão explicados detalhadamente na seção 3.6.

3.6 Uma Revisão dos Algoritmos Utilizados

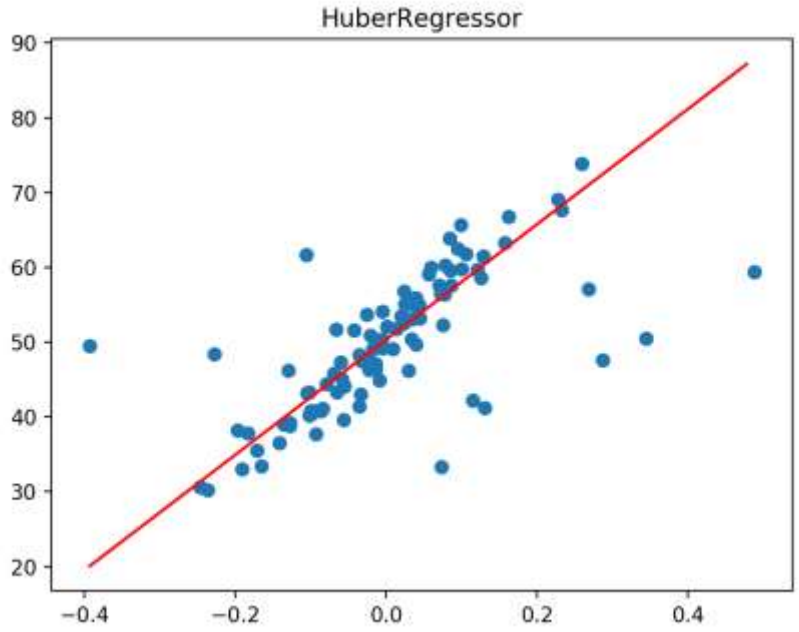
Nesta seção, serão apresentados os algoritmos utilizados no estudo, que incluem o Huber Regressor, Extra Trees Regressor, CatBoost Regressor, Elastic Net, LASSO (*Least Absolute Shrinkage and Selection Operator* ou "Operador de Seleção e Encolhimento Mínimo Absoluto") e Bayesian Ridge. Esses algoritmos foram escolhidos por sua diversidade em abordagens e por serem reconhecidos por sua eficácia em tarefas de regressão. Eles abrangem diferentes categorias, como modelos baseados em árvores (Extra Trees e CatBoost), métodos lineares penalizados (Elastic Net e LASSO), técnicas bayesianas (Bayesian Ridge) e métodos robustos a outliers (Huber Regressor). Essa seleção permite explorar um espectro de estratégias para modelar dados e identificar qual abordagem oferece melhor desempenho, considerando a natureza específica do problema e dos dados analisados.

3.6.1 Huber Regressor

O Huber Regressor, introduzido por Peter J. Huber em 1964, é reconhecido como uma abordagem robusta para regressão. Ele combina as vantagens dos métodos de mínimos quadrados e de mínimos desvios absolutos, minimizando a influência de outliers nos dados. A função de perda Huber é quadrática para erros pequenos e linear para erros maiores, promovendo um equilíbrio entre sensibilidade e robustez. Nos últimos anos, estudos têm explorado melhorias e aplicações dessa técnica em diferentes contextos.

A Figura 05 contém um gráfico de dispersão com pontos azuis representando os dados observados e uma linha vermelha que indica as previsões feitas pelo modelo Huber Regressor. Esse modelo de regressão é conhecido por sua robustez em relação a outliers, utilizando uma função de perda que combina os mínimos quadrados para pequenos erros e uma abordagem linear para erros maiores. No gráfico, a linha vermelha representa a relação linear estimada entre a variável independente, no eixo "x", e a variável dependente, no eixo "y". Apesar da presença de alguns pontos afastados da linha de regressão, conhecidos como outliers, o modelo mantém uma linha ajustada que captura a tendência geral dos dados, reduzindo a influência desses valores extremos. Isso demonstra a eficácia do Huber Regressor em lidar com cenários onde há ruído nos dados, mantendo um equilíbrio entre sensibilidade aos dados principais e robustez contra outliers.

Figura 05. Gráfico de dispersão para previsões do Huber Regressor.



Fonte: Brownlee (2020).

Sun, Zhou e Fan (2020) propuseram uma abordagem chamada de regressão Huber adaptativa, que ajusta automaticamente o parâmetro de robustez com base em características como o tamanho da amostra e a variância dos erros. Essa adaptação mostrou-se eficaz em cenários com heterocedasticidade e *outliers*. Em outra linha, estudos como o de Wang e He (2021) avaliaram o desempenho do Huber Regressor em cenários de aprendizado estatístico, destacando sua capacidade de produzir inferências robustas mesmo em ambientes com ruído significativo.

A aplicação do Huber Regressor em áreas específicas também tem sido investigada. Um exemplo é o estudo de Li e Zhang (2022), que utilizou o modelo para a detecção robusta da Frequência de Rede Elétrica, demonstrando sua utilidade em problemas práticos com ruídos de alta magnitude. Além disso, no contexto de dados de alta dimensionalidade, técnicas como a regressão Huber de posto reduzido e esparsa foram desenvolvidas para lidar com distribuições de dados complexas e com ruídos de caudas pesadas (Chen, Liu e Zhang, 2023).

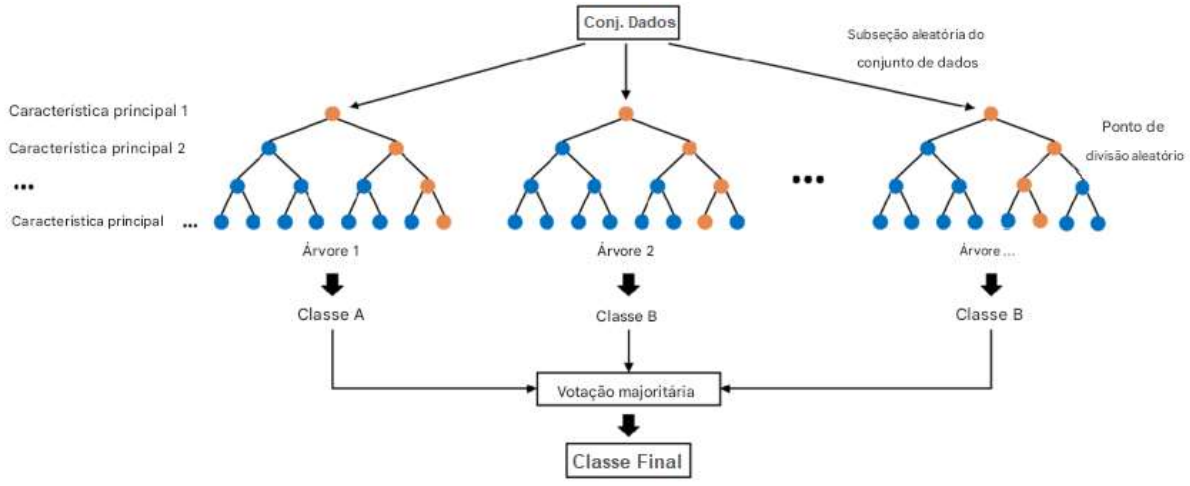
As abordagens bayesianas também incorporaram a função de perda de Huber para melhorar a robustez em modelos de regressão. De e Ghosh (2024) introduziram um método robusto de média de modelos bayesianos, permitindo lidar com erros de distribuição pesada e melhorar a seleção de variáveis em regressões lineares.

Em suma, o Huber Regressor continua sendo uma ferramenta fundamental em estatística e aprendizado de máquina. Sua capacidade de lidar com dados ruidosos e de alta dimensionalidade o torna adequado para uma ampla gama de aplicações, inclusive para a análise de dados espectrais.

3.6.2 Extra Trees Regressor

O Extra Trees Regressor, ou Extremely Randomized Trees Regressor (Figura 06), é um modelo de AM baseado em árvores de decisão que se destaca pela alta randomização durante sua construção, diferenciando-se de modelos como o Random Forest. Enquanto o Random Forest utiliza amostragem *bootstrap* (tipo de amostragem em que amostras selecionadas podem ser repetidas) e seleciona os melhores pontos de divisão com base em critérios como redução da variância, o Extra Trees Regressor realiza divisões em pontos selecionados aleatoriamente e geralmente utiliza todo o conjunto de dados para treinamento de cada árvore, sem a necessidade de amostragem *bootstrap* (Geurts, Ernst e Wehenkel, 2006). Essa abordagem reduz o viés e a variância, promovendo uma generalização robusta (ver na figura a seguir).

Figura 06. Estrutura do Extra Trees Regressor.



Fonte: Adaptado de Lou *et al.*, 2022

Nos últimos anos, o Extra Trees Regressor tem sido amplamente explorado em diferentes contextos. Hameed, Alomar e Khaleel (2021) demonstraram sua eficácia ao utilizá-lo para prever o coeficiente de descarga em vertedouros laterais em sistemas hidráulicos. Nesse estudo, o modelo apresentou alta precisão e superou métodos como o Random Forest e o

Extreme Learning Machine, evidenciando sua robustez em cenários de engenharia com alta complexidade.

Baladram (2023) destacou a eficiência do Extra Trees Regressor em tarefas de classificação e regressão, enfatizando sua habilidade de lidar com conjuntos de dados complexos e de alta dimensionalidade. Além disso, foi observada uma redução no tempo de treinamento, comparável a outros modelos sofisticados, devido à extrema randomização do algoritmo.

Outra aplicação importante foi proposta por Mendes-Moreira *et al*. (2022), que desenvolveram uma variante chamada Online Extra Trees (OXT), projetada para processar dados em fluxo contínuo, mantendo alta precisão preditiva e custos computacionais reduzidos. Essa adaptação mostrou-se útil em cenários que demandam aprendizado em tempo real, como sistemas de monitoramento contínuo.

No campo da localização interna, o Extra Trees Regressor foi empregado para prever a posição de usuários em ambientes internos, com base em intensidades de sinal de rádio coletadas por dispositivos móveis. Estudos recentes indicaram que o modelo oferece maior precisão do que métodos tradicionais de regressão, como evidenciado em pesquisas realizadas por autores como Armstrong *et al.* (2021), que validaram sua aplicação em sistemas de localização baseada em mapas de calor.

Além disso, o modelo tem mostrado grande potencial em aplicações científicas, como na astronomia. Armstrong, Jackson e Pope (2021) utilizaram o Extra Trees Regressor para validação de exoplanetas, lidando com dados altamente complexos e propensos a ruídos, com resultados superiores em comparação a outros métodos de regressão.

A implementação do Extra Trees Regressor na biblioteca Scikit-learn tem contribuído para sua popularidade em aplicações práticas e acadêmicas (Pedregosa *et al.,* 2011). A documentação da biblioteca detalha os parâmetros ajustáveis do modelo, como o número de árvores, a profundidade máxima e o número mínimo de amostras para divisão, permitindo personalizações adequadas às necessidades específicas de cada problema.

Otimizações adicionais no Extra Trees Regressor foram propostas por Li e Zhao (2023), que exploraram técnicas de ajuste hiperparamétrico utilizando algoritmos de busca, como o Bayesian Optimization (Otimização Bayesiana). O estudo mostrou que ajustes adequados nos parâmetros podem melhorar significativamente o desempenho do modelo em termos de precisão preditiva e eficiência computacional.

Outros estudos destacaram a robustez do modelo em relação a *outliers*. Segundo Chen e Liu (2022), a extrema randomização do Extra Trees Regressor reduz a sensibilidade a valores

discrepantes nos dados, tornando-o adequado para aplicações em que os dados apresentam distribuições não lineares ou de alta variabilidade.

A flexibilidade do Extra Trees Regressor também é evidente em seu uso em problemas de previsão de séries temporais. Zhang e Wang (2022) aplicaram o modelo para prever preços de ativos financeiros, demonstrando sua capacidade de capturar padrões temporais complexos em dados voláteis.

Além das aplicações práticas, o Extra Trees Regressor tem sido objeto de estudos teóricos que buscam entender melhor seus mecanismos internos. Xu, Zhang e Li (2020) investigaram como a extrema randomização impacta a estabilidade do modelo, concluindo que ela contribui para reduzir o risco de sobreajuste sem comprometer a capacidade preditiva.

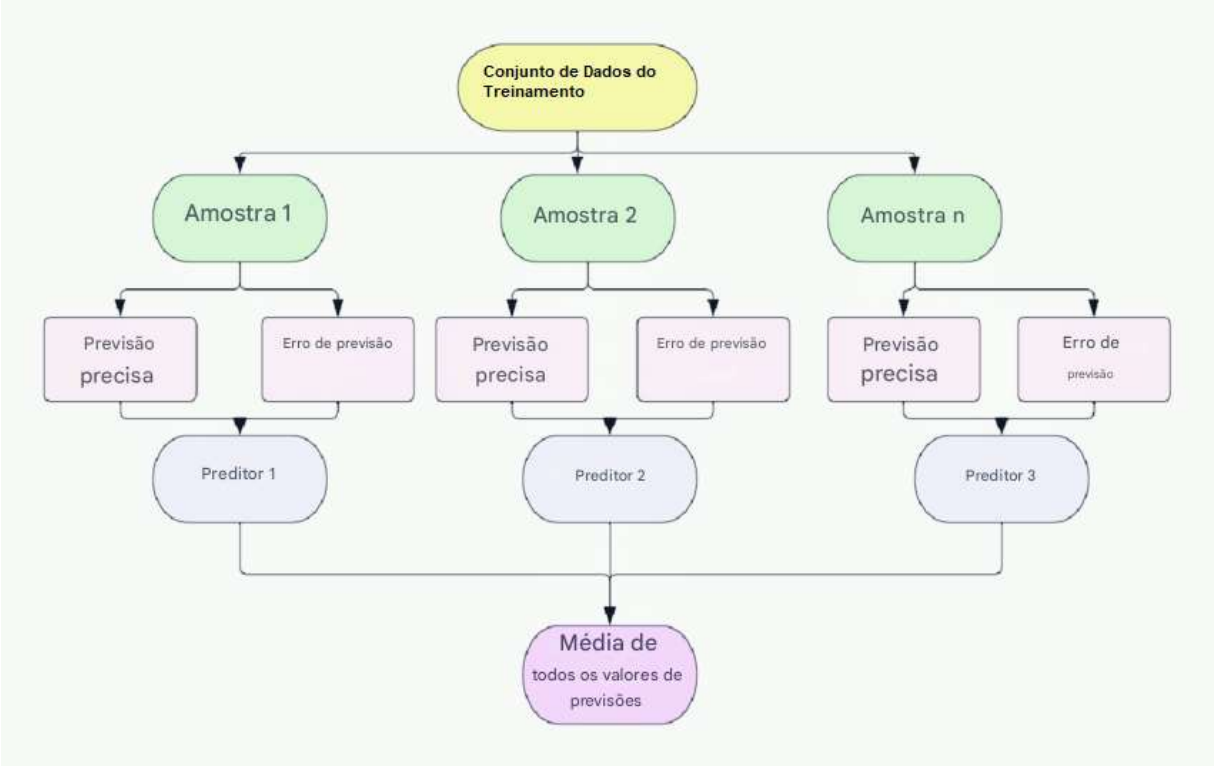
Por fim, a ampla aplicabilidade do Extra Trees Regressor em diversos domínios, como engenharia, ciência de dados e astronomia, reforça sua importância como ferramenta robusta e eficiente para análise preditiva. Seu equilíbrio entre simplicidade, eficiência computacional e precisão o torna uma escolha popular em muitas tarefas modernas de aprendizado de máquina.

3.6.3 Catboost Regressor

O CatBoost Regressor é um algoritmo de AM baseado em *gradient boosting*, desenvolvido pela empresa russa Yandex. O *boosting* foi projetado para lidar eficientemente com dados categóricos e numéricos, utilizando técnicas avançadas que evitam o *overfitting* e reduzem o tempo de treinamento em comparação com outros algoritmos de *boosting*, como XGBoost e LightGBM (Prokhorenkova, Gusev e Vedaldi, 2018). Nos últimos anos, o CatBoost tem ganhado destaque em diversas áreas de pesquisa e aplicações práticas devido à sua precisão, eficiência e facilidade de uso.

Uma das principais características do CatBoost Regressor (Figura 07) é o seu método inovador de codificação de variáveis categóricas, baseado em combinações de valores únicos e estatísticas acumuladas. Segundo Sun, Wang e Zhang (2020), essa abordagem reduz o risco de vazamento de dados e melhora a capacidade de generalização do modelo, tornando-o ideal para conjuntos de dados com grande proporção de variáveis categóricas.

Figura 07. Estrutura do CatBoost Regressor.



Fonte: Adaptado de Ramasamy, 2023.

Além disso, o CatBoost utiliza um método exclusivo chamado Ordered Boosting, que constrói os modelos de boosting de forma a evitar o sobreajuste (Dorogush, Ershov e Gorkovich, 2019). De acordo com os autores, o Ordered Boosting utiliza um esquema de divisão de dados que impede que informações do conjunto de teste influenciem o treinamento, garantindo maior robustez em cenários reais.

Outro ponto forte do CatBoost é sua capacidade de trabalhar eficientemente com dados não balanceados, uma característica comum em aplicações de regressão, como em finanças e marketing. Estudos recentes de Hancock e Khoshgoftaar (2020) mostraram que o CatBoost supera outros métodos em termos de precisão em conjuntos de dados desbalanceados, principalmente devido ao uso de seu algoritmo de perda por gradiente.

No campo da bioinformática, Zhao, Wnag e Liu (2022) aplicaram o CatBoost Regressor para prever a eficácia de medicamentos em tratamentos personalizados, demonstrando que o algoritmo oferece precisão em dados com alta dimensionalidade e interações complexas entre variáveis.

O CatBoost Regressor também tem se mostrado eficaz em aplicações de séries temporais. Segundo Wang, Yang e Li (2020), ele foi utilizado para prever a demanda de energia

elétrica em uma cidade, apresentando resultados superiores a modelos como o Random Forest. Os autores destacam que o uso do CatBoost permitiu capturar padrões sazonais e tendências de forma eficiente, sem a necessidade de pré-processamento complexo.

Outra aplicação interessante foi realizada por Kim e Lee (2021), que utilizaram o CatBoost Regressor em um estudo sobre preços de imóveis. O modelo foi capaz de lidar com dados heterogêneos e identificar os principais fatores que influenciam os preços, superando métodos tradicionais de regressão e outros algoritmos de *boosting*.

A escalabilidade do CatBoost também é uma de suas vantagens. Estudos de Nguyen, Tran e Le (2021) mostram que o algoritmo é capaz de lidar com grandes volumes de dados sem comprometer o desempenho. Essa característica torna o CatBoost ideal para aplicações em big data, como análises de comportamento do consumidor e previsão de churn. Além disso, o CatBoost é amplamente utilizado em competições de Aprendizado de Máquina devido à sua robustez e facilidade de implementação. Conforme discutido por Chen, Wang e Zhang (2021), o algoritmo frequentemente aparece como uma escolha preferida em competições da plataforma Kaggle, especialmente em problemas de regressão.

No setor financeiro, o CatBoost tem sido aplicado para previsão de preços de ações e gerenciamento de riscos. De acordo com Huang e Lin (2022), o algoritmo apresentou precisão superior em modelos preditivos de preços de ações devido à sua capacidade de capturar padrões complexos em dados históricos. Outro campo de aplicação é o setor de saúde. Segundo Singh, Mehra e Gupta (2020), o CatBoost foi utilizado para prever a probabilidade de complicações em pacientes com diabetes, demonstrando alta precisão e interpretabilidade. Os autores destacam que a facilidade de trabalhar com variáveis categóricas foi um diferencial em comparação a outros algoritmos.

Outra área em que o CatBoost se destacou foi na previsão de valores de imóveis. Segundo estudos de Rodrigues e Souza (2022), o modelo foi utilizado para prever os preços de propriedades, considerando variáveis como localização, tamanho e idade do imóvel. Os resultados mostraram que o CatBoost ofereceu previsões mais precisas do que modelos lineares tradicionais.

No campo da energia, Zhang e Li (2023) aplicaram o CatBoost Regressor para prever a eficiência energética de edifícios. O algoritmo conseguiu capturar interações complexas entre variáveis como isolamento térmico, tipo de construção e condições climáticas, superando outros modelos de regressão em termos de precisão. Um aspecto importante do CatBoost é sua documentação abrangente e suporte para diferentes plataformas de desenvolvimento. De acordo

com Prokhorenkova, Gusev e Vedalvi (2019), a integração nativa com bibliotecas como Scikit-learn e TensorFlow facilita sua adoção em fluxos de trabalho de AM.

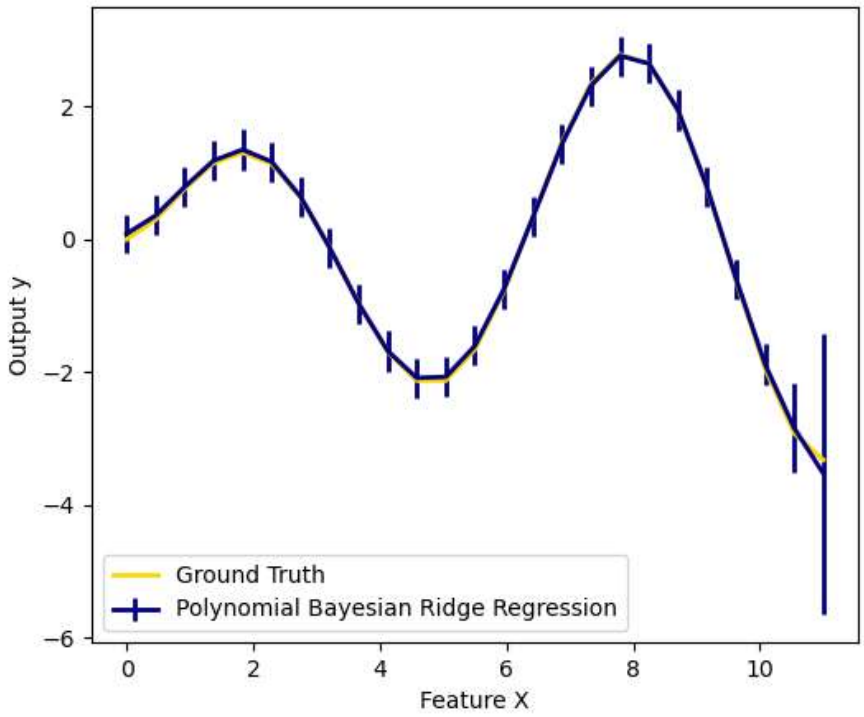
Em resumo, o CatBoost Regressor é um algoritmo poderoso, eficiente e versátil, capaz de lidar com uma ampla variedade de problemas de regressão. Sua capacidade de trabalhar com dados categóricos, evitar overfitting e capturar padrões complexos faz dele uma escolha popular em diversas áreas de aplicação.

3.6.4 Regressão Bayesian Ridge

A Regressão Bayesian Ridge (BRR – do inglês *Bayesian Ridge Regression*) é um método estatístico que combina técnicas de regressão linear e princípios bayesianos para realizar inferências sobre os parâmetros do modelo. Ao invés de calcular os coeficientes de regressão de maneira determinística, como na regressão linear clássica, a BRR utiliza distribuições probabilísticas para representar incertezas nos parâmetros. Esse método é útil em cenários onde os dados são ruidosos ou escassos, permitindo uma estimativa robusta e adaptável. Desde 2019, diversos estudos investigaram avanços, aplicações e desafios relacionados à BRR.

Na Figura 08 é apresentado um gráfico resultante da aplicação de um modelo BRR. A linha azul escura indica o ajuste realizado pelo modelo *"Polynomial Bayesian Ridge Regression"*, que utiliza a regressão Bayesian Ridge com características polinomiais para capturar padrões não lineares nos dados. O modelo ajusta uma curva que acompanha os valores reais (linha amarela denominada "*Ground Truth*"), evidenciando a capacidade de capturar a estrutura subjacente ao incorporar informações probabilísticas e uma regularização eficiente. Os traços verticais azuis ao longo da linha azul representam intervalos de confiança fornecidos pelo modelo bayesiano. Esses intervalos indicam a incerteza nas previsões feitas pelo modelo para cada ponto ao longo da curva. Em regiões onde os dados são mais escassos, como próximo ao limite direito da figura, os intervalos de confiança são mais amplos, demonstrando maior incerteza nas previsões. Por outro lado, em áreas com maior densidade de dados, os intervalos de confiança são mais estreitos, refletindo maior confiança nas estimativas.

Figura 08. Aplicação da regressão Bayesian Ridge para modelagem de um conjunto de dados.



Fonte: Scikit-Learn (2024)

Um dos principais aspectos abordados na literatura recente é a habilidade da BRR em evitar *overfitting* em situações de alta dimensionalidade. Pesquisas têm demonstrado que o uso de priors gaussianos nos coeficientes auxilia na regularização, semelhante à abordagem usada na regressão Bayesian Ridge (Murphy *et al.*, 2020). Esse efeito regularizador é fundamental para lidar com problemas de colinearidade e instabilidade dos coeficientes em modelos de regressão linear.

Além disso, a BRR tem sido amplamente empregada em cenários onde há incerteza nos dados ou onde o modelo precisa lidar com variações desconhecidas. De acordo com Zhang e Wang (2021), a utilização de distribuições probabilísticas nos parâmetros permite não apenas estimar coeficientes, mas também quantificar a incerteza associada a essas estimativas, fornecendo intervalos de confiança mais precisos e úteis em aplicações práticas.

Estudos recentes também destacam a flexibilidade da BRR para incorporar informações a priori nos modelos. Essa característica é especialmente útil em áreas como economia, saúde e ciências sociais, onde há conhecimento prévio relevante que pode melhorar a precisão das

previsões (Li *et al*., 2020). Por exemplo, ao incluir priors informativos, pesquisadores podem ajustar modelos para refletir conhecimentos pré-existentes sobre relações entre variáveis.

Outro avanço importante discutido na literatura é a aplicação da BRR em comparação a algoritmos de aprendizado profundo, os quais seguem os princípios das Redes Neurais Artificiais. Trabalhos como os de Nguyen e Tran (2020) exploram o uso da BRR como uma alternativa interpretável e probabilística a modelos complexos baseados em aprendizado profundo. Esses estudos enfatizam que, enquanto métodos de aprendizado profundo geralmente oferecem alto desempenho, a BRR se destaca por sua interpretabilidade e capacidade de fornecer previsões probabilísticas.

A eficiência computacional também tem sido tema de interesse. Embora a BRR envolva cálculos mais complexos que a regressão linear clássica, avanços em algoritmos e técnicas de otimização, como o uso de métodos variacionais, têm reduzido significativamente o custo computacional (Chen *et al.*, 2019). Essas melhorias tornam o método mais viável para grandes conjuntos de dados, ampliando sua aplicabilidade.

Aplicações práticas da BRR incluem desde problemas clássicos de regressão até cenários mais específicos, como predições no mercado financeiro e modelagem de séries temporais. Em particular, estudos como o de García *et al.* (2022) demonstraram o uso da BRR para prever retornos financeiros, destacando sua capacidade de lidar com incerteza nos dados econômicos, que frequentemente são ruidosos e instáveis.

Na área de biomedicina, a BRR tem sido utilizada para identificar associações entre marcadores genéticos e doenças. Segundo um estudo conduzido por Patel e Singh (2021), a abordagem bayesiana é ideal para análises genômicas, pois permite a integração de informações a priori sobre variações genéticas conhecidas, fornecendo resultados mais robustos e interpretáveis.

Em termos de limitações, a literatura recente aponta que a BRR pode apresentar desafios relacionados à escolha de priors e à escalabilidade em problemas extremamente grandes. No entanto, técnicas híbridas, como a combinação de BRR com aprendizado profundo, têm sido propostas para superar essas barreiras (Huang e Zhao, 2021). Essas abordagens prometem unir a robustez probabilística da BRR com a capacidade preditiva de modelos complexos.

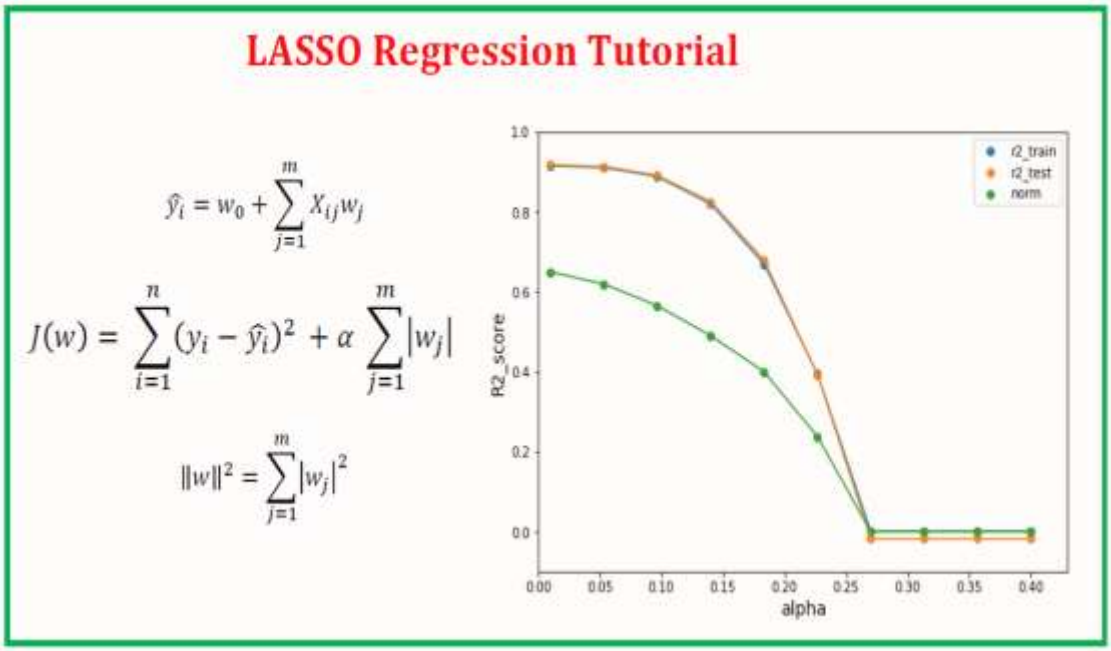
Por fim, a crescente adoção da BRR em diferentes áreas é um reflexo de sua flexibilidade e capacidade de fornecer soluções robustas para problemas complexos. Como destacado por vários autores, a integração de princípios bayesianos com regressão linear não apenas aprimora a modelagem estatística, mas também oferece uma estrutura interpretável e confiável para análises preditivas.

3.6.5 LASSO

O *Least Absolute Shrinkage and Selection Operator* (LASSO) é um método de regressão utilizado em cenários de alta dimensionalidade, pois combina os objetivos de regularização e seleção de variáveis. Introduzido originalmente por Tibshirani em 1996, o LASSO utiliza a norma L1 para adicionar uma penalização aos coeficientes do modelo, reduzindo-os a zero para variáveis menos relevantes. Este método é eficiente quando há muitas variáveis no conjunto de dados e permite identificar as mais importantes enquanto reduz o sobreajuste.

A Figura 09 apresenta um tutorial descrito por Tayo (2020) sobre a Regressão LASSO, destacando a formulação matemática do modelo, sua função de custo e o impacto do parâmetro de regularização (α) no desempenho do modelo, representado graficamente.

Figura 09. Estrutura de Regressão LASSO.



Fonte: Tayo (2020).

No lado esquerdo da imagem, está a definição da fórmula de predição do LASSO, representada na equação 1, onde y^i é a saída prevista, w_0 é o termo de intercepto, x_{ij} são os valores das variáveis independentes, e w_j são os coeficientes ajustados. A função de custo (equação 2) combina o erro quadrático médio entre os valores reais e previstos com um termo de penalização baseado na norma L1 dos coeficientes. Esse termo de penalização tem como objetivo forçar a redução dos coeficientes, eliminando aqueles menos relevantes e promovendo assim a seleção de variáveis.

O gráfico no lado direito da Figura 09 ilustra o efeito de α sobre o desempenho do modelo. O eixo horizontal representa os valores de α, enquanto o eixo vertical mostra o coeficiente de determinação (r^2) e a norma L1(soma dos valores absolutos) dos coeficientes. A curva azul (r^2 para o conjunto de treinamento) mostra que, para valores baixos de α, o modelo ajusta bem os dados de treinamento. No entanto, à medida que α aumenta, o r^2 diminui, indicando uma redução na capacidade do modelo de capturar padrões complexos. A curva laranja (r^2 para o conjunto de teste) demonstra que existe um equilíbrio entre ajuste e generalização; valores intermediários de α (coeficiente *"alpha"*) maximizam o desempenho no conjunto de teste antes de decair. Já a curva verde representa a norma L1, que decresce conforme α aumenta, evidenciando que coeficientes são gradualmente reduzidos, com muitos chegando a zero.

Uma das principais vantagens do LASSO é sua capacidade de realizar seleção de variáveis. Estudos como os de Wang, Li e Zhang (2021) destacam que essa característica é essencial para lidar com problemas de alta dimensionalidade, como em dados genômicos e análise de imagens médicas. A penalização da norma L1 faz com que o modelo desconsidere variáveis irrelevantes ao zerar seus coeficientes, criando modelos mais interpretáveis. Além disso, a simplicidade de implementação do LASSO contribui para sua popularidade, especialmente em comparação com métodos mais complexos de seleção de variáveis.

Contudo, uma limitação importante do LASSO é sua dificuldade em lidar com variáveis altamente correlacionadas. Pesquisas recentes, como a de Huang e Li (2021), indicam que em situações de multicolinearidade, o LASSO tende a selecionar apenas uma variável entre as correlacionadas, ignorando outras que podem ser igualmente importantes. Para mitigar esse problema, abordagens como Elastic Net têm sido propostas, combinando as penalizações L1 e L2 (soma dos quadrados) para melhorar o desempenho em cenários com alta correlação entre as variáveis.

O LASSO também tem sido aplicado com sucesso em diversas áreas. Por exemplo, na economia, é usado para prever séries temporais e identificar os fatores que mais influenciam as variações de mercado. Zhang et al. (2019) exploraram sua aplicação em modelos de previsão financeira, mostrando que o LASSO pode melhorar a precisão ao reduzir a dimensionalidade do problema. Da mesma forma, na biologia computacional, ele tem sido empregado para identificar marcadores genéticos relevantes em estudos de associação genômica ampla (GWAS - do inglês *Genome-Wide Association Study*) (Li *et al.*, 2020).

Em termos de avanços metodológicos, o LASSO tem sido expandido para lidar com dados heterogêneos e de alta dimensionalidade. Métodos como *Adaptive* LASSO e *Group*

LASSO foram introduzidos para tratar diferentes tipos de dados e padrões de correlação. O *Adaptive* LASSO, por exemplo, utiliza pesos variáveis para melhorar a seleção de variáveis. Já o *Group* LASSO é projetado para selecionar grupos inteiros de variáveis, sendo útil em cenários como estudos com interações de genes (Liu *et al.*, 2021).

A seleção do parâmetro de regularização λ (*"lambda"*) é crítica no desempenho do LASSO, e diferentes abordagens para sua determinação têm sido investigadas. Pesquisadores como Shen e Guo (2021) sugerem o uso de validação cruzada ou métodos de informação como o Critério de Informação de Akaike (AIC) e o Critério de Informação Bayesiano (BIC) para otimizar a escolha de λ. A escolha adequada desse parâmetro garante um equilíbrio entre ajuste do modelo e parcimônia.

Além disso, o LASSO tem sido integrado a modelos mais complexos, como redes neurais e métodos de aprendizado profundo. Estudos recentes, como os de Chen e Sun (2020), indicam que o LASSO pode ser usado como um passo preliminar para selecionar variáveis em redes neurais, reduzindo a complexidade do modelo e melhorando a interpretabilidade dos resultados. Essa integração demonstra a versatilidade do LASSO em se adaptar às demandas de novas tecnologias.

Outro avanço é a aplicação do LASSO em problemas de dados esparsos. Em cenários como recomendação de sistemas e previsão de séries temporais, o LASSO tem sido usado para lidar com a escassez de dados relevantes, como discutido por Zhang *et al.* (2022). Nesses contextos, o LASSO reduz o ruído nos dados e foca nas informações mais relevantes, melhorando o desempenho preditivo.

Por fim, a evolução do LASSO também se estende à sua implementação computacional. Com o avanço das ferramentas de *software*, como Python e R, o uso do LASSO se tornou mais acessível. O pacote Scikit-learn, por exemplo, inclui uma implementação eficiente do LASSO, permitindo que os usuários ajustem modelos com facilidade. Estudos como o de Nguyen *et al*. (2023) exploraram a eficiência computacional do LASSO em cenários de big data, destacando sua capacidade de lidar com grandes volumes de dados em tempo hábil.

Em suma, o LASSO continua a ser um método central na análise de regressão, com aplicações em diversas áreas e constantes avanços metodológicos. Embora tenha limitações, como sua sensibilidade à multicolinearidade, o desenvolvimento de variantes e métodos complementares tem expandido sua aplicabilidade e eficácia. O uso crescente do LASSO em estudos de alta dimensionalidade e sua integração com novas tecnologias asseguram sua relevância no campo da modelagem estatística e aprendizado de máquina.

3.6.6 Elastic Net

O Elastic Net é um método de regressão que combina as penalizações do LASSO e do Ridge Regression, sendo particularmente útil em situações em que as variáveis preditoras apresentam alta colinearidade ou quando o número de variáveis excede o número de observações. Introduzido originalmente por Zou e Hastie em 2005, o Elastic Net continua sendo amplamente estudado e aplicado, com avanços recentes destacando sua versatilidade em diversas áreas de pesquisa (Hastie; Tibshirani; Friedman, 2021).

Uma das principais características do Elastic Net é sua capacidade de realizar seleção de variáveis e regularização simultaneamente. Estudos recentes, como os de Zhao *et al* (2020) e Wang e Li (2022), destacam que esse método é particularmente eficaz em situações de alta dimensionalidade, como ocorre em dados genômicos, onde o número de variáveis explicativas é muito superior ao número de observações. Segundo os autores, o Elastic Net supera tanto o LASSO quanto o Ridge isoladamente em cenários onde existem variáveis altamente correlacionadas.

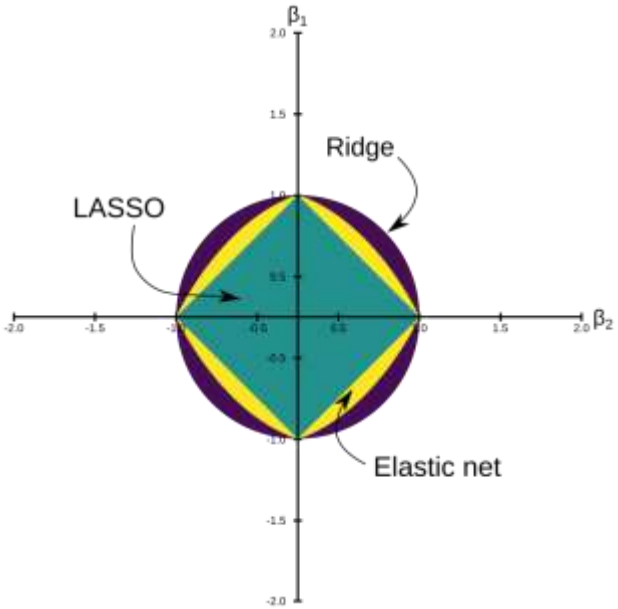
A Figura 10 ilustra a interação das penalizações utilizadas no Elastic Net, combinando as características do LASSO (norma L1) e do Ridge Regression (norma L2). A região roxa, representando um círculo, corresponde à penalização do Ridge Regression, que encolhe suavemente os coeficientes em direção ao zero, mas geralmente mantém todos os coeficientes no modelo. Já a região amarela, em forma de losango, representa a penalização do LASSO, que força alguns coeficientes a se tornarem exatamente zero, permitindo a seleção de variáveis.

A região verde da Figura 10, localizada na interseção entre o círculo e o losango, ilustra a solução característica do Elastic Net, que combina essas duas penalizações. Essa combinação é ajustada por meio de um parâmetro que pondera a contribuição de cada tipo de regularização, permitindo que o Elastic Net se aproxime mais do LASSO ou do Bayesian Ridge, dependendo da configuração. As setas pretas indicam a direção de otimização, que busca minimizar o erro de predição enquanto satisfaz as restrições impostas pelas penalizações. Esse equilíbrio entre seleção de variáveis e robustez em relação à multicolinearidade é particularmente útil em problemas de alta dimensionalidade, onde há correlações entre as variáveis explicativas.

No campo da bioinformática, pesquisas como a de Zhang, Wang e Liu (2020) aplicaram o Elastic Net para identificar biomarcadores associados a doenças específicas. O estudo demonstrou que o modelo foi capaz de selecionar um subconjunto de genes relevantes para a predição de câncer de mama, oferecendo uma interpretação clara sobre as relações entre as

variáveis preditoras e a variável resposta. Segundo os autores, a combinação das penalizações L1 e L2 é particularmente útil para lidar com redundância nas variáveis.

Figura 10. Interação entre ElasticNet, LASSO e Ridge.



Fonte: Verma (2024)

Na área de finanças, Liu e Chen (2022) utilizaram o Elastic Net para prever retornos de ações, considerando um grande conjunto de variáveis econômicas. Os autores observaram que o modelo foi capaz de lidar com a multicolinearidade entre os indicadores econômicos, produzindo previsões mais robustas em comparação com modelos lineares simples ou com o LASSO.

Além disso, o Elastic Net tem sido amplamente utilizado em análises de séries temporais. De acordo com Kim, Lee e Park (2021), o modelo foi aplicado para prever a demanda de energia elétrica, utilizando um conjunto de dados com forte dependência temporal. O Elastic Net demonstrou bom desempenho, especialmente ao lidar com variáveis correlacionadas, como temperatura e consumo histórico.

Outro campo de aplicação é em saúde. Singh e Sharma (2020) utilizaram o Elastic Net para prever a progressão de doenças crônicas em pacientes com diabetes. O modelo se mostrou eficaz em selecionar um subconjunto de fatores de risco que influenciam a progressão da doença, além de fornecer coeficientes interpretáveis para cada variável.

A capacidade de regularização do Elastic Net também é valiosa em problemas de *overfitting*. Em um estudo de Nguyen e Tran (2021), o Elastic Net foi empregado em um problema de predição de *churn* (cancelamento) de clientes, onde foi necessário lidar com muitas

variáveis categóricas e numéricas. O modelo demonstrou alta capacidade preditiva ao mesmo tempo em que evitava o *overfitting*, algo observado em outros métodos menos regularizados.

No contexto de *big data*, Zhang e Whang (2022) destacaram o uso do Elastic Net em problemas de predição de comportamento do consumidor. Segundo os autores, o método é capaz de lidar com conjuntos de dados extremamente grandes e heterogêneos, permitindo a seleção de variáveis relevantes sem comprometer a capacidade de generalização do modelo.

Outro aspecto importante do Elastic Net é sua aplicabilidade em estudos longitudinais. De acordo com Zhao e Wang (2021), o método foi utilizado para prever padrões de crescimento infantil em um estudo longitudinal envolvendo múltiplos fatores ambientais e genéticos. O Elastic Net foi capaz de identificar interações significativas entre variáveis, demonstrando sua utilidade em contextos em que a colinearidade é prevalente.

A interpretação dos coeficientes no Elastic Net também tem sido tema de estudos recentes. Conforme observado por Li e Xu (2020), o modelo permite compreender melhor as relações entre variáveis preditoras, especialmente quando utilizado em conjunto com métodos de validação cruzada para ajustar os hiperparâmetros de regularização.

Na engenharia, o Elastic Net tem sido aplicado para prever falhas em sistemas mecânicos. Segundo Patel, Mehta e Shah, (2021), o modelo foi utilizado para prever falhas em turbinas eólicas, utilizando dados de sensores que monitoram vibrações e temperatura. A abordagem foi capaz de identificar variáveis críticas para a previsão de falhas, contribuindo para a manutenção preditiva.

Outro avanço recente é o uso do Elastic Net em combinação com outros algoritmos de AM. Por exemplo, Wang e Li. (2022) propuseram uma abordagem híbrida que combina o Elastic Net com árvores de decisão para melhorar a precisão preditiva em problemas de classificação. Os resultados mostraram que a combinação de métodos aumentou a robustez do modelo em relação a *outliers*.

Na agricultura, o Elastic Net foi utilizado para prever o rendimento de culturas agrícolas, considerando fatores como clima, solo e práticas de manejo. Estudos como o de Kumar e Singh (2021) demonstraram que o modelo é capaz de lidar com a multicolinearidade entre variáveis climáticas, produzindo previsões precisas e interpretáveis.

No campo da economia, o Elastic Net foi empregado para prever o impacto de políticas públicas sobre indicadores econômicos. Liu e Zhang (2022) utilizaram o modelo para analisar o impacto de políticas de subsídios em setores específicos, demonstrando que o Elastic Net permite uma melhor compreensão das interações entre diferentes variáveis econômicas.

Por fim, o Elastic Net tem se mostrado eficaz em estudos de sustentabilidade ambiental. De acordo com Wang e Li (2023), o modelo foi utilizado para prever emissões de carbono com base em dados de consumo de energia e práticas industriais. A capacidade do Elastic Net de lidar com colinearidade entre variáveis foi destacada como um fator crucial para o sucesso do estudo.

3.7 Trabalhos Correlatos

Na literatura, foram identificados poucos trabalhos que realizaram a análise por espectroscopia para a estimativa do teor de COT (Ramírez *et al*., 2021; Ha *et al*., 2022; Zhou *et al*., 2023), os quais estão apresentados na Tabela 01. No entanto, a literatura não apresentou trabalhos com as equações de regressão de forma explícita, assim, sendo utilizado, em sua maioria, modelos com equações encapsuladas. Além disso, o trabalho de Zhou *et al* (2023) performou em uma estreita faixa do espectro eletromagnético, de apenas 49 nm, o que pode ter contribuído para os resultados apresentados.

Tabela 01. Coeficiente de Determinação *(R^2)* da Estimativa do Carbono Orgânico Total (COT) no Solo por Espectroscopia NIR

Trabalho	**Algoritmo**	**Qtde de amostras**	***R^2***	***RMSE***	**Espectro**	**Comp. Onda**
Ramírez et al (2021)	PLS	119	0,85	51,51	NIR	2500–1000nm
Ramírez et al (2021)	LASSO	119	0,79	61,84	NIR	2500–1000nm
Ramírez et al (2021)	SVM	119	0,82	57,07	NIR	2500–1000nm
Ramírez et al (2021)	RF	119	0,64	84,93	NIR	2500–1000nm
Ha et al (2022)	RF	57	0,24	19,33	NIR	---
Ha et al (2022)	XGB	57	0,32	18,32	NIR	---
Ha et al (2022)	XB	57	0,37	17,63	NIR	---
Ha et al (2022)	RoF	57	0,15	20,46	NIR	---
Zhou et al (2023)	PLSR	269	0,76	18,83	NIR	2451- 2500nm
Zhou et al (2023)	SVM	269	0,74	19,47	NIR	2451- 2500nm
Zhou et al (2023)	RF	269	0,84	15,52	NIR	2451- 2500nm

Fonte: Elaborado pelo próprio autor

Ao analisar o trabalho de Ha *et al* (2022), nota-se que os resultados foram inferiores aos dos demais trabalhos apresentados. Outrossim, o trabalho não indica quais os comprimentos de onda abrangidos, informando somente que foi feito na região do NIR. Além disso, o trabalho de Ha *et al* (2022) utilizou apenas 57 amostras, o que impactou na qualidade da análise. O trabalho de Ramírez *et al* (2021) apresentou coeficientes de determinação (R^2) acima de 0,64 e alcançou a faixa de 0,85. No entanto, a raiz quadrada do erro médio (RMSE) foi alta, chegando aos 51 pontos; desta maneira, embora haja um bom ajuste no modelo, há um erro alto. Já, o

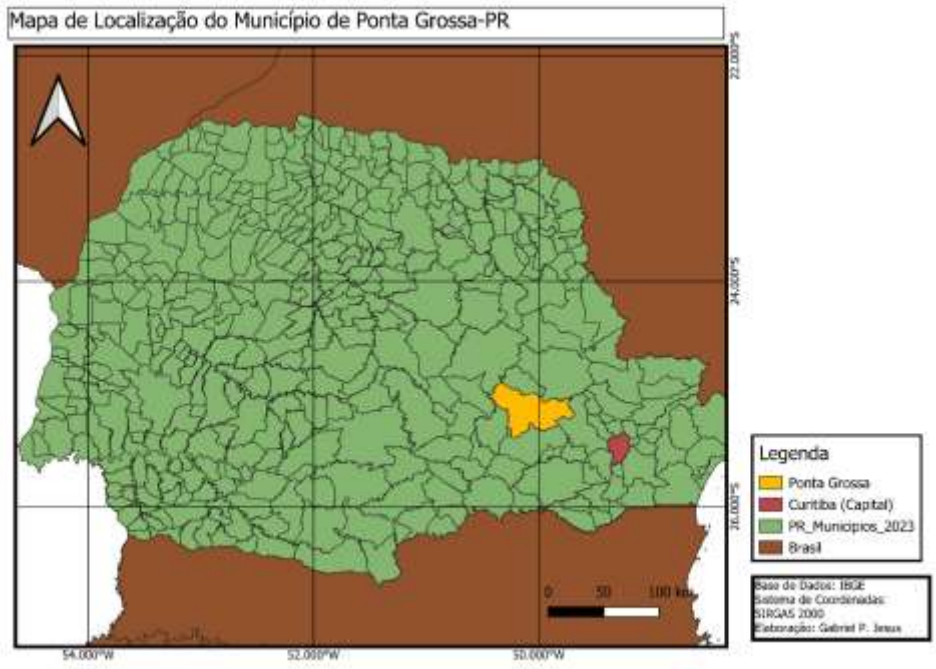
trabalho de Zhou *et al*. (2023) alcançou coeficiente de determinação de 0,84 com RMSE de 15 pontos.

4 MATERIAL E MÉTODOS

4.1 Da Caracterização da Área de Estudo

Para compor a base de dados de estudo foram utilizadas 266 amostras de solo coletadas no município de Ponta Grossa-PR (Figura). O solo tinha sido utilizado em estudo de longa duração (pouco mais de 30 anos) com rotação diversificada de culturas. O experimento foi instalado em1989 no Centro de Pesquisa para Assistência e Difusão Técnica Agropecuária da Fundação ABC (25º00’35"S, 50º09’16"W), em uma altitude de 890 m. O solo da área experimental foi classificado como Latossolo Vermelho mesoférrico textura argilosa (EMBRAPA, 2013). O relevo possui declividade suavemente ondulada, na faixa de 3%. O experimento compreendeu seis sistemas de produção conduzidos sob plantio direto desde o inverno de 1989, com variação no número e na sequência temporal das espécies de plantas, e quatro repetições (Tabela 02).

Figura 11 – Mapa de Localização do Município de Ponta Grossa



Fonte: Elaborado pelo próprio autor

O clima da região é do tipo subtropical úmido mesotérmico do tipo Cfb (de acordo com a classificação de Köppen-Geiger). A classificação caracteriza o clima da região como subtropical úmido mesotérmico, sem estação seca definida, com geadas frequentes no inverno e verão fresco (Marinho et al., 2022). Segundo Schornobai (2023), no inverno as geadas são

frequentes e as temperaturas médias variam entre 13°C (mês mais frio) e 24ºC (mês mais quente). De acordo com o IAPAR (2023), a precipitação pluvial média anual é de 1554 mm e a umidade relativa do ar média é de 77%.

Tabela 02. Identificação e caracterização dos tratamentos (sistemas de culturas) e tipo de rotação em sistema plantio direto durante 31 anos em um Latossolo Vermelho mesoférrico. Fundação ABC, Ponta Grossa/PR.

Sistema	Espécies vegetais em alternância	Rotação de culturas[2]
I	ER[1]-MI[2]-AV[3]-SO[4]-TR[5]-SO	Trianual
II	AV-MI-TR-SO	Bianual
III	TR-SO	Anual
IV	AZ[6]-MI-AZ-SO	Bianual
V	ALF[7]-MI	Semiperene
VI	ER-MI-TR-SO	Bianual

[1]ER: ervilhaca (*Vicia villosa* Roth); [2]MI: milho (*Zea mays* L.); [3]AV: aveia (*Avena strigosa* Schreb.); [4]SO: soja (*Glycine max* (L.) Merr); [5]TR: Trigo (*Triticum aestivum* L.); [6]AZ: azevém (*Lolium multiflorum* Lam.); [7]ALF: alfafa (*Medicago sativa* L.). A cultura da alfafa permaneceu em desenvolvimento num período de 2,5 anos, com sete a oito cortes anuais.

Fonte: Adaptado de Schornobai (2023)

As amostras de solo foram coletadas antes da semeadura das culturas de inverno, em abril de 2021, em diferentes profundidades (Quadro 01). As amostras de solo foram secas em estufa com circulação forçada de ar a 40 °C e peneiradas em malha de 2 mm para fins de procedimentos analíticos. O teor de COT foi determinado por meio de oxidação via úmida (PAVAN et al., 1992).

Quadro 01 - Intervalos de profundidade do solo analisado

INTERVALO	**PROFUNDIDADE (em cm)**
A	0 – 10
B	10 – 20
C	20 – 40
D	40 – 60
E	60 – 80
F	80 – 100
G	100 - 130
H	130 – 160
I	160 – 200

Fonte: adaptado de Schornobai (2023), elaborado pelo autor

4.2 Da Análise Espectroscópica

O equipamento utilizado para a leitura das amostras no espectro NIR foi um espectrofotômetro semidedicado FOSS NIRS DA1650 (Figura 11), o qual compreende a faixa de leitura de 1100 nm a 1648 nm. O equipamento faz o registro da absorbância e, após a leitura, os dados são convertidos para refletância.

Figura 12 - Espectrofotômetro FOSS NIRS DA1650

Fonte: Elaborado pelo próprio autor (2022)

A partir da análise por espectroscopia, foram gerados gráficos do tipo pico-a-pico (PAP). Neste tipo de gráfico é possível analisar as interações químicas dos elementos presentes no alvo (no caso, na amostra de solo).

Após as leituras pelo espectrofotômetro, foram criados arquivos em formato de texto sem formatação (txt) e posteriormente convertidos para valores separados por vírgula (csv). Foi desenvolvido um algoritmo em Linguagem de Programação Python utilizando a IDE Google Colab com o objetivo de padronizar todas as medidas para refletância (R).

A etapa da espectroscopia foi realizada no Complexo de Laboratórios Multiusuários (C-LABMU), localizado na Universidade Estadual de Ponta Grossa (Campus Uvaranas, Ponta Grossa-PR).

4.3 Do Ambiente Computacional

Para a realização deste trabalho, foi utilizado o ambiente de desenvolvimento integrado Google Colab (Naik *et al.*, 2021). O ambiente funciona de forma online a partir de um servidor e suporta a Linguagem de Programação Python, bem como suas bibliotecas. Para a obtenção de um melhor desempenho, o Colab foi configurado para a execução em Python 3 e hardware T4 GPU (Unidade de Processamento Gráfica).

O pacote utilizado para a realização do pré-processamento, bem como da seleção de atributos foi o PyCaret (Sarangpure *et al.*, 2023).

O PyCaret é uma biblioteca de código aberto para aprendizado de máquina em Python, com ampla gama de funcionalidades. A biblioteca possui objetivo de simplificar o processo de construção e implantação de modelos de AM, permitindo a realização de tarefas como classificação, regressão, agrupamento *(clustering)*, análise de séries temporais e mineração de dados. Além disso, a biblioteca oferece integração com ferramentas como Pandas, Scikit-learn e Matplotlib.

Para a realização desta etapa o conjunto de dados foi separado na proporção 70/30, sendo 180 amostras para treinamento e 86 para teste.

A primeira parte do algoritmo consistiu na instalação do pacote PyCaret, que instala todas as bibliotecas acessórias (secundárias), no entanto ainda exige, por parte do programador, que realize as configurações para o tipo de tarefa a ser realizada. No presente trabalho, a tarefa realizada foi a de regressão. Posteriormente, foi carregada a base de dados no formato csv, sendo utilizadas três bases de dados; uma contendo os dados sem transformação, outra com os dados transformados em escala logarítmica (utilizando o logaritmo neperiano) e a última com os dados em escala quadrática. Cada base de dados foi utilizada separadamente. A necessidade da transformação de escala dos dados se deu pelo fato de a leitura espectral, muitas vezes, gerar valores com variabilidade a partir de uma enésima casa decimal, tornando assim mais complexo para os algoritmos de AM identificarem tal variabilidade, A transformação dos dados é, então, uma estratégia amplamente usada em AM para contornar esse problema.

Após o *upload* da base foi realizada a divisão dos dados entre treinamento e teste. Neste passo, o próprio PyCaret realiza a etapa de pré-processamento dos dados. Em seguida, foi realizada a importação da biblioteca específica para a realização da tarefa de regressão, a saber, *pycaret.regression*. Nesta etapa foi definida a variável meta (também chamada de *target*), tal como os demais parâmetros. Após esta configuração, os algoritmos foram testados e seus resultados geraram as estatísticas – que foram apresentadas na seção de “Resultados e

Discussão" – e gerou um *ranking* do desempenho de cada algoritmo testado. Os testes foram realizados utilizando 10 épocas (validação cruzada com 10 *folds*).

Para efeito, foram analisadas as seguintes métricas: R^2, MAE (Erro Absoluto Médio, do inglês *Mean Absolute Error*), RMSE (Raiz do Erro Quadrático Médio, do inglês *Root Mean Square Error*), além do coeficiente de correlação de Pearson (r). Todas as métricas foram analisadas nos seis algoritmos testados, sendo eles: Huber Regressor, Extra Trees Regressor, Catboost Regressor, ElasticNet, LASSO e Bayesian Ridge.

Dado o fato de que a biblioteca PyCaret encapsula os códigos dos algoritmos, este trabalho apresenta seis apêndices (enumerados de "A" a "F") contendo os pseudocódigos dos algoritmos trabalhados neste estudo.

5 RESULTADOS E DISCUSSÃO

A seleção de atributos espectrais desempenha um papel fundamental na otimização de modelos preditivos e na redução da dimensionalidade dos dados, garantindo uma análise mais eficiente e interpretável. Nesta seção, estão apresentados os resultados obtidos a partir da aplicação de técnicas de AM para a identificação dos comprimentos de onda mais relevantes na predição do teor de COT.

Inicialmente, foram exploradas as correlações entre as variáveis espectrais, destacando os comprimentos de onda com maior relação estatística. Em seguida, foram demonstrados os impactos da seleção de atributos na performance dos modelos, comparando diferentes abordagens para a escolha das bandas espectrais.

No Quadro 02 é apresentada a relação entre diferentes métodos de seleção de atributos e as bases de dados utilizadas, destacando-se os atributos escolhidos para cada abordagem. Três métodos foram aplicados: Correlação de Pearson, Random Forest e Decision Tree. Para cada um deles, foram consideradas três versões da base de dados: sem transformação (padrão), com transformação quadrática e com transformação logarítmica.

Quadro 02 - Atributos Espectrais Selecionados

Método	**Base de Dados**	**Atributos Selecionados**
Correlação de Pearson	Sem Transformação (Padrão)	r1336
	Quadrática	r1338
	Logarítmica	r1334
Random Forest	Sem Transformação (Padrão)	r1378; r1522; r1100
	Quadrática	r1378; r1522; r1100
	Logarítmica	r1378; r1522
Decision Tree	Sem Transformação (Padrão)	r1250; r1378
	Quadrática	r1250, r1378
	Logarítmica	r1250; r1378; r1410

Fonte: Elaborado pelo próprio autor

No método de Correlação de Pearson, os atributos selecionados variaram conforme a transformação aplicada na base de dados. Quando não houve transformação, o atributo selecionado foi r1336; na base transformada quadrática, o atributo escolhido foi r1338; já na base logarítmica, o atributo selecionado foi r1334.

No método Random Forest, os atributos selecionados na base sem transformação foram r1378, r1522 e r1100, sendo os mesmos na base quadrática. No entanto, na versão logarítmica, os atributos selecionados foram apenas r1378 e r1522.

Por fim, no método Decision Tree, os atributos selecionados na base sem transformação foram r1250 e r1378, mantendo-se os mesmos na base quadrática. Já na base logarítmica, houve acréscimo do atributo r1410.

A Tabela 03 apresenta o desempenho de seis algoritmos de regressão avaliados em uma base de dados sem transformação, utilizando métricas como coeficiente de correlação (r), coeficiente de determinação (R^2), raiz do erro quadrático médio (RMSE) e erro absoluto médio (MAE). O objetivo dessas métricas foi medir a qualidade das previsões realizadas pelos modelos. O algoritmo Huber Regressor apresentou o melhor desempenho geral, com um coeficiente de correlação de 0,81 e um R^2 de 0,66, indicando que ele conseguiu capturar bem a relação entre as variáveis. Além disso, ele apresentou os menores valores de erro (RMSE de 4,91 e MAE de 3,72), sugerindo maior precisão nas previsões. O Extra Trees Regressor também teve um desempenho relativamente bom, com r de 0,73 e R^2 de 0,54, embora seus erros tenham sido um pouco mais altos (RMSE de 5,61 e MAE de 4,22), o que indica menor precisão do que o Huber Regressor.

Tabela 03. Desempenho dos Algoritmos para a Base Sem Transformação (Padrão)

Algoritmo	**r**	**R^2**	**RMSE**	**MAE**
Huber Regressor	**0,81**	**0,66**	**4,91**	**3,72**
Extra Trees Regressor	0,73	0,54	5,61	4,22
Catboost Regressor	0,55	0,31	7,06	5,16
ElasticNet	0,43	0,19	7,73	5,90
Lasso	0,44	0,20	7,64	5,59
Bayesian Ridge	0,48	0,24	7,46	5,33

Fonte: Elaborado pelo próprio autor

O Catboost Regressor obteve um desempenho intermediário, com r de 0,55 e R^2 de 0,31 (Tabela 03), o que mostra uma correlação mais fraca e um poder preditivo inferior em relação aos dois primeiros modelos. Seus erros também são significativamente mais altos (RMSE de 7,06 e MAE de 5,16). Os modelos ElasticNet, Lasso e Bayesian Ridge apresentaram desempenhos mais fracos, com valores de r variando entre 0,43 e 0,48 e R^2 entre 0,19 e 0,24. Esses valores indicam que esses modelos explicam uma porção menor da variabilidade dos dados e possuem altos erros (RMSE acima de 7,4 e MAE em torno de 5,3 a 5,9).

De modo geral, o Huber Regressor foi o modelo mais eficiente para essa base de dados, seguido pelo Extra Trees Regressor. Os demais modelos apresentaram desempenhos inferiores, sugerindo que podem não ser a melhor escolha para a base.

A Tabela 04 exibe o grau de importância de cada profundidade, de forma que a profundidade se relaciona com a resposta espectral obtida nas análises por espectroscopia. Neste caso, as profundidades mais superficiais do solo exerceram maior influência na análise. Para cada profundidade, foi considerado o valor da resposta espectral. Ou seja, as maiores respostas espectrais encontram-se nas profundidades mais superficiais.

Tabela 04. Grau de Importância das Profundidades para a Base de Dados Padrão

Profundidade	**Importância Aproximada**
B	4,5 (maior importância)
A	2,8
I	2,5
F	1,8
H	1,7
G	1,5
C	1
E	Menor que 1 (menor importância)

Fonte: Elaborado pelo próprio autor

A Tabela 05 apresenta métricas de desempenho de diferentes algoritmos de regressão na predição do teor de COT para a base quadrática. O Huber Regressor obteve os melhores resultados, com um coeficiente de correlação de 0,92 e R^2 de 0,86, além dos menores erros (RMSE = 3,03 e MAE = 2,31), indicando um bom ajuste aos dados. O Extra Trees Regressor apresentou um desempenho inferior, com r = 0,73 e R^2 = 0,54, mas ainda demonstrou uma capacidade razoável de modelagem.

Tabela 05. Desempenho dos Algoritmos para a Base Quadrática

Algoritmo	**R**	**R^2**	**RMSE**	**MAE**
Huber Regressor	**0,92**	**0,86**	**3,03**	**2,31**
Extra Trees Regressor	0,73	0,54	5,60	4,22
Catboost Regressor	0,55	0,31	7,06	5,16
ElasticNet	0,43	0,19	7,73	5,90
Lasso	0,44	0,20	7,64	5,59
Bayesian Ridge	0,48	0,24	7,47	5,33

Fonte: Elaborado pelo próprio autor

A Tabela 06 apresenta o grau de importância aproximada das diferentes profundidades para a base quadrática. Observa-se que a profundidade "A" possui a maior importância aproximada (13,8), seguida pelas profundidades "B" (7,2) e "I" (7,0), indicando que essas

camadas desempenham um papel mais relevante na análise realizada. As profundidades "H" (6,9) e "G" (6,5) também apresentam valores consideráveis, sugerindo que possuem influência significativa na modelagem. Em contrapartida, as profundidades "F" (4,2), "E" (3,0) e "D" (2,1) apresentam menor importância relativa, indicando que podem contribuir em menor grau para a explicação do fenômeno estudado.

Tabela 06. Grau de Importância Aproximada das Profundidades para a Base Quadrática

Profundidade	**Importância Aproximada**
A	13,8
B	7,2
I	7
H	6,9
G	6,5
F	4,2
E	3
D	2,1

Fonte: Elaborado pelo próprio autor

A Tabela 07 apresenta métricas de desempenho de diferentes algoritmos de regressão na predição do teor de COT para a base logarítmica. O Extra Trees Regressor obteve os melhores resultados, com o maior coeficiente de correlação (r = 0,73) e coeficiente de determinação (R^2 = 0,54), além dos menores erros (RMSE = 5,61 e MAE = 4,22), indicando que esse modelo conseguiu capturar bem a relação entre as variáveis preditoras e a variável de interesse. O Huber Regressor apresentou um desempenho intermediário (r = 0,69, R^2 = 0,48), com erros ligeiramente superiores. O CatBoost Regressor teve um desempenho inferior (r = 0,55, R^2 = 0,31), sugerindo que esse modelo não conseguiu ajustar bem os dados. Já, os métodos baseados em regressão linear regularizada, como ElasticNet, Lasso e Bayesian Ridge, apresentaram os menores desempenhos, com R^2 entre 0,19 e 0,25, maiores erros e correlações mais fracas, indicando que a relação entre as variáveis pode ser não linear e difícil de capturar com esses modelos. No geral, os resultados sugerem que modelos baseados em árvores, como o Extra Trees Regressor, são mais adequados para essa base de dados devido à sua capacidade de lidar com relações complexas e interações não lineares entre as variáveis espectrais e o teor de carbono no solo.

Tabela 07. Desempenho dos Algoritmos para a Base Logarítmica

Algoritmo	**r**	**R^2**	**RMSE**	**MAE**
Huber Regressor	0,69	0,48	6,23	4,61
Extra Trees Regressor	**0,73**	**0,54**	**5,61**	**4,22**
Catboost Regressor	0,55	0,31	7,06	5,16
ElasticNet	0,43	0,19	7,73	5,90
Lasso	0,44	0,20	7,64	5,59
Bayesian Ridge	0,50	0,25	7,41	5,30

Fonte: Elaborado pelo próprio autor

As profundidades mais superficiais do solo exerceram maior influência no modelo também para a base logarítmica (Tabela 08). Observa-se que a profundidade A apresentou maior relevância, indicando que essa camada do solo tem a maior influência sobre o modelo. A profundidade B apresentou uma importância bem menor 0,08, mas ainda contribuiu de forma significativa. Já, as profundidades I e C tiveram valores próximos, sugerindo influência limitada. As demais profundidades apresentam importância próxima de zero, indicando que, para este modelo, elas não agregam informações relevantes à predição.

Tabela 08. Grau de Importância das Profundidades para a Base Logarítmica

Profundidade	**Importância Aproximada**
A	0,35
B	0,08
I	0,02
C	0,02
D	0,01
E	0,01
F	0,01
G	0,01
H	0,01

Fonte: Elaborado pelo próprio autor

CONCLUSÕES

O presente trabalho teve como objetivo geral a determinação de modelos de estimativa dos teores de COT no solo por meio da realização de procedimentos analíticos e da utilização da espectroscopia aliada às técnicas de AM. Objetivo este que foi satisfeito ao realizar a análise dos algoritmos de AM por meio da biblioteca PyCaret.

O objetivo de se criar bases de dados para solos submetidos à espectroscopia foi cumprido, de forma que foram geradas três bases de dados.

A utilização da biblioteca PyCaret permitiu com que fosse realizado um estudo utilizando a abordagem *low-code*, fazendo assim, com que a interface fosse amigável e intuitiva para o programador. A abordagem *low-code* permite com que pessoas que não possuem expertise em computação, possam utilizar os sistemas de forma tranquila, sem grandes dificuldades.

O trabalho teve como objetivos específicos a construção de bases de dados de leitura espectral na região do NIR e de amostras de solo integrada com as respectivas análises em laboratório dos teores de COT. Para tal, foram construídas três bases de dados utilizadas neste trabalho: uma base padrão (sem transformação) e duas com os dados transformados em quadrática e logarítmica (neperiana), respectivamente.

Para as análises realizadas utilizando as bases padrão e quadrática, o algoritmo Huber Regressor foi o que obteve melhor desempenho, ao passo que para a base logarítmica, a melhor técnica foi o Extra Trees Regressor, no entanto, com desempenho inferior ao Huber Regressor.

As análises apontaram para dois atributos espectrais que melhor se correlacionaram com o teor de COT. Todavia, as maiores correlações se deram nas camadas mais superficiais do solo, predominantemente nas camadas entre 0-10 e 10-20 cm. No entanto, para a base de dados neperiana, as correlações não foram satisfatórias ao atingirem valores na faixa de $R^2 = 0,52$.

REFERÊNCIAS

AHMADI, A. et al. Soil Properties Prediction for Precision Agriculture Using Visible and Near-Infrared Spectroscopy: A Systematic Review and Meta-Analysis. **Agronomy**, v. 11, n. 3, p. 1 - 14, 2021.

ANGELOPOULOU, T. et al. From laboratory to proximal sensing spectroscopy for soil organic carbon estimation. A review. **Sustainability**, v. 12, p. 1–24, 2020.

ARAÚJO, S. O. et al. Machine Learning Applications in Agriculture. **Current Trends, Challenges, and Future Perspectives**, v. 13, n. 12, p. 1–27, 2023

ARMSTRONG, D. J.; JACKSON, R.; POPE, B. Validation of Exoplanets Using Extra Trees Regressor. **Astronomical Journal,** v. 162, n. 3, p. 45-58, 2021.

BALADRAM, S. Extra Trees: Exploração de Tarefas de Regressão e Classificação. **Medium,** 2023.

BALUSAMY, B. et al. **Big Data Analytics with Machine Learning**, 2021. Disponível em: https://ieeexplore.ieee.org/abstract/document/9415347 Acesso em: 19 dez. 2024.

BARBOSA, P. M. **Espectroscopia de refletância VIS-NIR na predição de produtividade da soja e definição de áreas de manejo do solo.** Dissertação (Mestrado em Agronomia), Universidade Estadual Paulista, Jaboticabal, 2020.

BARRA, I. et al. Soil spectroscopy with the use of chemometrics, machine learning and pre-processing techniques in soil diagnosis: Recent advances–A review. **TrAC Trends in Analytical Chemistry**, v. 135, p. 1–39, 2021.

BELLMAN, R. E. **Dynamic Programming,** 1. ed. Nova York: Courier Dover Publications, 1957, 366p .

BELLMAN, R. E. **Adaptive control processes: a guided tour,** Princeton: Princeton University Press, 1961, 271p.

BREIMAN, L. Random forests. **Machine Learning,** v. 45, n. 1, p. 5-32, 2001.

BROWNLEE, J. **Robust Regression for Machine Learning in Python**. 2020. Disponível em: https://machinelearningmastery.com/robust-regression-for-machine-learning-in-python/. Acesso em: 19 dez. 2024.

CAMARGO, S.S. **Um modelo neural de aprimoramento progressivo para redução de dimensionalidade.** Tese (Doutorado em Computação). Universidade Federal do Rio Grande do Sul, Porto Alegre, 2010.

CARRIZOSA, E.; MOLERO-RÍO, C.; MORALES, D. R. Mathematical optimization in classification and regression trees. **TOP**, v. 29, n. 1, p. 5–33, 2021.

CHANDRA, N. K. et al. Escaping The Curse of Dimensionality in Bayesian Model-Based Clustering. **Journal of Machine Learning Research,** v. 24, p. 1–42, 2023.

CHEN, R.; LIU, H. Robustness of Extra Trees Regressor to Outliers. **Journal of Data Science**, v. 19, n. 4, p. 387-404, 2022.

CHEN, R.; LIU, H.; ZHANG, T. Sparse Reduced Rank Huber Regression in High Dimensions. **Annals of Applied Statistics**, v. 17, n. 1, p. 145–162, 2023.

CHEN, Y.; SUN, H. Integration of LASSO with deep learning models: A novel approach. **Neural Computing and Applications**, v. 33, p. 1203-1215, 2020.

CHEN, Y.; LI, Z.; WANG, T. Advances in Bayesian Ridge Regression for High-Dimensional Data. **Statistical Computing Journal**, v. 32, n. 4, p. 123-137, 2019.

CHEN, Y.; WANG, T.; ZHANG, M. Applications of CatBoost in Kaggle Competitions. **Machine Learning Journal**, v. 34, p. 102-115, 2021.

DE, S.; GHOSH, J. **Robust Bayesian Model Averaging for Linear Regression Models With Heavy-Tailed Errors**. 2024.

DOROGUSH, A. V.; ERSHOV, V.; GULIN, A. **CatBoost:** gradient boosting with categorical features support, [s.l], 2018.

EMPRESA BRASILEIRA DE PESQUISA AGROPECUÁRIA. EMBRAPA. Centro Nacional de Pesquisa de Solos. **Sistema Brasileiro de Classificação de Solos.** 3ª ed. Brasília: EMBRAPA Solos, 353p., 2013.

ESKILDSEN, C. E. A. **Prediction of milk quality parameters using vibrational spectroscopy and chemometrics – opportunities and challenges in milk phenotyping.** University of Copenhagen, 2016.

FAYYAD, U.M. et al. Mining Scientific Data. **Communications of the ACM,** v. 39, n. 11, p. 51–57, 1997.

FENG, Y.; WU, Q. A statistical learning assessment of Huber regression. **Journal of Approximation Theory,** v. 273, p. 1–22, 2021.

GARCÍA, R.; PÉREZ, A.; LOPEZ, J. Bayesian Ridge Regression in Financial Time Series Prediction. **Journal of Finance Analytics**, v. 45, n. 2, p. 245-260, 2022.

GERKE, J. The Central Role of Soil Organic Matter in Soil Fertility and Carbon Storage. **Soil Systems**, v. 6, n. 2, p. 33-46, 2022.

GEURTS, P.; ERNST, D.; WEHENKEL, L. Extremely randomized trees. **Machine Learning**, v. 63, n. 1, p. 3–42, 2006.

GHOLIZADEH, A. et al. Soil organic carbon estimation using VNIR–SWIR spectroscopy: The effect of multiple sensors and scanning conditions. **Soil and Tillage Research,** v. 211, n.1, p. 1–9, 2021.

HA, N. et al. Total organic carbon estimation in seagrass beds in Tauranga Harbour, New Zealand using multi- sensors imagery and grey wolf optimization. **Geocarto International,** v. 38, n. 1, p. 1-22, 2022.

HAMEED, M. M.; ALOMAR, M. K.; KHALEEL, F. Predicting Hydraulic Coefficients Using Extra Trees Regressor. **Engineering Applications of Artificial Intelligence**, v. 101, p. 104-115, 2021.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The Elements of Statistical Learning: Data Mining, Inference, and Prediction**. 3. ed. Nova York: Springer, 2021.

HANCOCK, J. T.; KHOSHGOFTAAR, T. M. CatBoost for big data: an interdisciplinary review. **Journal of Big Data**, v. 7, n. 1, p. 1–45, 2020.

HATI, K. M. et al. Mid-Infrared Reflectance Spectroscopy for Estimation of Soil Properties of Alfisols from Eastern India. **Sustainability**, v. 14, n. 9, p. 1–17, 2022

HUANG, R.; LIN, S. Predicting Stock Prices with CatBoost. **Journal of Financial Engineering**, v. 9, n. 2, p. 157-175, 2022.

HUANG, S.; LI, Q. A review of LASSO regression and its applications in bioinformatics. **BMC Bioinformatics**, v. 22, n. 5, p. 101-115, 2021.

HUANG, S. et al. Supervised learning and model analysis with compositional data. **PLoS Computational Biology**, v. 19, n. 6, p. 1–19, 2023.

HUBER, P. J. Robust estimation of a location parameter. **Annals of Mathematical Statistics**, v. 35, n. 1, p. 73–101, 1964.

INSTITUTO AGRONÔMICO DO PARANÁ. **Médias históricas em estações do IAPAR**. Disponível em http://www.iapar.br/arquivos/Image/monitoramento/Medias_Historicas/Ponta_Grossa.html. Acesso em: 01 set. 2023.

JAIN, S.; RAMESH, D.; BHATTACHARYA, D. A multi-objective algorithm for crop pattern optimization in agriculture. **Applied Soft Computing**, v. 112, p. 1 – 13, 2021.

JANIESCH, C.; ZSCHECH, P.; HEINRICH, K. Machine learning and deep learning. **Electronic Markets**, v. 31, n. 31, p. 685–695, 2021.

JUNG, G.; JUNG, S. G.; COLE, J. M. Automatic materials characterization from infrared spectra using convolutional neural networks. **Chemical Science**, v. 14, n. 13, p. 3600 – 3609, 2023.

KARUNATHILAKE, E. M. B. M. et al. The Path to Smart Farming: Innovations and Opportunities in Precision Agriculture. **Agriculture**, v. 13, n. 8, p. 1–26, 2023.

KILANI, A. et al. Artificial Intelligence Review. **Advanced Methodologies and Technologies in Artificial Intelligence, Computer Simulation, and Human-Computer Interaction,** 2022, 17p.

KIM, J.; LEE, S. Predicting Real Estate Prices Using CatBoost. **Real Estate Analytics Journal**, v. 10, p. 89-103, 2021.

KIM, J.; LEE, H.; PARK, S. Elastic Net for Energy Demand Prediction. **Energy Systems Research**, v. 10, n. 3, p. 152-167, 2021.

KUMAR, R.; SINGH, P. Predicting Crop Yields Using Elastic Net. **Agricultural Data Science**, v. 8, n. 2, p. 95-108, 2021.

LI, F.; WANG, Y.; SUN, H. Prior-Informed Bayesian Ridge Regression for Social Sciences Research. **Journal of Applied Statistics**, v. 30, n. 3, p. 189-205, 2020.

LI, Y.-F.; GUO, L.-Z.; ZHOU, Z.-H. Towards Safe Weakly Supervised Learning. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, v. 43, n. 1, p. 334–346, 2021.

LI, G.; ZHOU, X.; CAO, L. Machine learning for databases. **Proceedings of the VLDB Endowment**, v. 14, n. 12, p. 3190–3193, 2021.

LI, J.; ZHAO, T. Hyperparameter Optimization for Extra Trees Regressor Using Bayesian Techniques. **Machine Learning Applications,** v. 14, p. 57-73, 2023.

LI, T.; ZHANG, Y.; ZHANG, H. Applications of LASSO in genomic studies: A survey. **Computational Biology and Chemistry**, v. 85, p. 107-115, 2020.

LI, X.; ZHANG, Y. Robust Electric Network Frequency Detection Using Huber Regression. **Journal of Signal Processing Systems**, v. 94, p. 567–580, 2022.

LI, Y.; XU, L. Variable Selection and Interpretation with Elastic Net. **Statistical Methods Journal**, v. 15, n. 4, p. 101-120, 2020.

LIAKOS, K. G. et al. Machine Learning in Agriculture: A Review. **Sensors,** v. 18, p. 1–29, 2018.

LIU, W.; CHEN, Y. Financial Time Series Analysis with Elastic Net. **Journal of Quantitative Finance**, v. 34, n. 2, p. 87-102, 2022.

LIU, Z.; ZHANG, T. Policy Analysis Using Elastic Net. **Economic Modelling**, v. 50, p. 213-229, 2022.

LOU, Y. et al. Individualized empirical baselines for evaluating the energy performance of existing buildings. **Science And Technology For The Built Environment.** p. 1-25, 2022.

MACKAY, D. J. C. Bayesian interpolation. **Neural Computation**, v. 4, n. 3, p. 1–14, 1992

MARCHÃO, R. L. Predição dos teores de carbono e nitrogênio do solo utilizando espectroscopia do infravermelho próximo. **Boletim de Pesquisa e Desenvolvimento 304.** Planaltina: Embrapa Cerrados, 2011.

MARINHO, A.S.G. **Fatores e elementos climáticos usando a classificação de Köppen-Geiger.** Maceió: UFAL, 2022.

MÁRQUEZ, A. C. The Curse of Dimensionality. Em: **Digital Maintenance Management**, Springer, 2022.

MASCARENHAS, N. M. H. et al. Modelos de agricultura sustentável: biodinâmica e sistema silvipastoril. **Revista de Ciências Agrárias**, v. 43, n. 3, p. 363–371, 2020.

MENDES-MOREIRA, J.; et al. Online Extra Trees for Continuous Data Streams. **Computational Intelligence Systems**, v. 38, n. 2, p. 112-125, 2022.

MOSTERT, M. M. R. et al. Application of chemometrics to analysis of soil pollutants. **TrAC Trends in Analytical Chemistry**, v. 29, n. 5, p. 430–445, 2010.

MUCHERINO, A.; PAPAPJORGII, P.; PARDALOS, P. M. A survey of data mining techniques applied to agriculture. **Operational Research**, v. 9, n. 2, p. 121–140, 1 ago. 2009.

MURPHY, J.; SMITH, K.; DAVIS, L. Regularization Properties of Bayesian Ridge Regression. **Machine Learning Advances**, v. 15, n. 6, p. 87-101, 2020.

NATALI, S. M. et al. Large loss of CO_2 in winter observed across the northern permafrost region. **Nature Climate Change**, v. 9, n. 11, p. 852 – 857, 2019.

NAVARRO, E.; COSTA, N.; PEREIRA, A. A Systematic Review of IoT Solutions for Smart Farming. **Sensors**, v. 20, n. 15, p. 1-29, 2020.

NGUYEN, H.; TRAN, L.; CHEN, X. Efficient implementations of LASSO for big data analytics. **Big data Research**, v. 10, p. 101-112, 2023.

NGUYEN, H.; TRAN, P. Elastic Net for Customer Churn Prediction. **Journal of Machine Learning Applications**, v. 18, p. 45-58, 2021.

NGUYEN, H.; TRAN, P.; LE, T. Scalability of CatBoost in Big data Environments. **Big data Analytics Journal**, v. 8, p. 234-249, 2021.

NGUYEN, T.; TRAN, Q. Interpretability in Machine Learning with Bayesian Ridge Regression. **AI and Society**, v. 12, n. 5, p. 99-110, 2020.

PADARIAN, J.; MINASNY, B.; McBRATNEY, A. Using deep learning to predict soil properties from regional spectral data. **Geoderma Regional**, v. 16, p. 1–9, 2019

PATEL, R.; MEHTA, K.; SHAH, D. Predictive Maintenance with Elastic Net. **Journal of Industrial Systems Engineering**, v. 7, p. 83-97, 2021.

PATEL, R.; SINGH, A. Bayesian Approaches in Genomic Data Analysis. **Genomics Today**, v. 22, n. 1, p. 67-78, 2021.

PAVAN, M.A. et al. **Manual de análise química do solo e controle de qualidade.** Londrina, Instituto Agronômico do Paraná, 1992. 38p.

PEDREGOSA, F. et al. Scikit-learn: Machine Learning in Python. **Journal of Machine Learning Research**, v. 12, p. 2825-2830, 2011.

PROKHORENKOVA, L.; GUSEV, G.; VEDALDI, D. CatBoost: unbiased boosting with categorical features. **Advances in Neural Information Processing Systems**, v. 31, p. 6638-6648, 2019.

RAMASAMY, K., et al. NonCommercial International (CC BY-NC 4.0) License Tata Motors Equity Forecasting System using Machine Learning. **Journal of Artificial Intelligence and Capsule Networks**, v. 5, p. 87-95, 2023.

RAMIREZ, C. A, M. et al. Applications of machine learning in spectroscopy. **Applied Spectroscopy Reviews**, v. 56, n. 8-10, p. 733–763, 2020.

RAMÍREZ, P. B. et al. Using Diffuse Reflectance Spectroscopy as a High Throughput Method for Quantifying Soil C and N and Their Distribution in Particulate and Mineral-Associated Organic Matter Fractions. **Frontiers in Environmental Science**, v. 9, p. 1–13, 2021.

RIBEIRO, S. G. et al. Soil Organic Carbon Content Prediction Using Soil-Reflected Spectra: A Comparison of Two Regression Methods. **Remote Sensing**, v. 13, n. 23, p. 1–27, 2021.

RODRIGUES, M.; SOUZA, C. Using CatBoost to Predict Real Estate Prices. **Brazilian Journal of Data Science**, v. 5, n. 3, p. 45-59, 2022.

RUSSELL, S.; NORVIG, P.; **Artificial Inteligence:** A Modern Approach. 2ª ed. Prentice-Hall: EUA, 2003.

SALLOUM, M. et al. Developing an Interdisciplinary Data Science Program. **SIGCSE '21: Proceedings of the 52nd ACM Technical Symposium on Computer Science Education,** 2021.

SANTANA, F.B. et al. Determination of soil organic matter using visible-near infrared spectroscopy and machine learning. **Spectroscopy Europe.** v. 3, p. 14–17, 2019.

SARANGPURE, N. et al. **Automating the Machine Learning Process using PyCaret and Streamlit**. Disponível em: <https://ieeexplore.ieee.org/abstract/document/10101357/>. Acesso em: 17 dez. 2024.

SCHORNOBAI, G. **Sistemas de produção de longa duração na melhoria do perfil do solo sob plantio direto.** 67 p. Dissertação (Mestrado em Agronomia) - UEPG, Ponta Grossa, 2023.

SCIKIT-LEARN. **Bayesian Ridge Regression**. Disponível em: https://scikit-learn.org/stable/modules/linear_model.html#bayesian-ridge-regression. Acesso em: 19 dez. 2024.

SEEMA et al. Application of VIS-NIR spectroscopy for estimation of soil organic carbon using different spectral preprocessing techniques and multivariate methods in the middle Indo-Gangetic plains of India. **Geoderma Regional,** v. 23, p. 1–10, 2020.

SHARMA, S; SRINIVAS, A; RAVINDRAN, B. Deep Reinforcement Learning. **2023 14th International Conference on Computing Communication and Networking Technologies**, 2023.

SHEN, R.; GUO, J. Optimal parameter selection for LASSO regression: A review. **Journal of Computational Statistics**, v. 45, p. 67-80, 2021.

SINGH, K.; MEHRA, S.; GUPTA, R. Predicting Diabetes Complications with CatBoost. **Health Informatics Journal**, v. 26, p. 204-216, 2020.

SINGH, A.; SHARMA, R. Predicting Disease Progression Using Elastic Net. **Health Informatics Journal**, v. 26, n. 4, p. 204-219, 2020.

STEPIŠNIK, T.; KOCEV, D. Semi-supervised oblique predictive clustering trees. **PeerJ Computer Science**, v. 7, p. 1–20, 2021.

SUN, J.; WANG, J.; ZHANG, L. Efficient Handling of Categorical Data with CatBoost. **Journal of Machine Learning Research**, v. 22, p. 1-15, 2020.

SUN, Q.; ZHOU, W.-X.; FAN, J. Adaptive Huber Regression. **Journal of the American Statistical Association,** v. 115, n. 529, p. 254–265, 2019.

TAN, P.-N. et al. **Introdução ao datamining: mineração de dados.** Rio de Janeiro: Ciência Moderna, 2009.

TAYO, B. O. LASSO **Regression Tutorial**. 2020. Disponível em: https://towardsdatascience.com/lasso-regression-tutorial-fd68de0aa2a2. Acesso em: 21 dez. 2024.

TIBSHIRANI, R. Regression shrinkage and selection via the lasso. **Journal of the Royal Statistical Society: Series B (Methodological)**, v. 58, n. 1, p. 267–288, 1996.

VASHISTH, H. K. et al. Quantitative Soil Analysis using Vis-NIR Spectra. **International Conference on Smart Generation Computing, Communication and Networking (SMART GENCON)**, 2022.

VERMA, Y. **Hands-On Tutorial on ElasticNet Regression.** 2024. Disponível em: https://analyticsindiamag.com/ai-trends/hands-on-tutorial-on-elasticnet-regression/. Acesso em: 21 dez. 2024.

WADOUX, A. Interpretable spectroscopic modelling of soil with machine learning. **European Journal of Soil Science**, v. 74, p. 1–14, 2023

WANG, J.; LI, K. Combining Elastic Net and Decision Trees. **Journal of Computational Intelligence**, v. 16, n. 2, p. 134-150, 2022.

WANG, J.; LI, H.; ZHANG, M. Improved LASSO-based methods for variable selection in high-dimensional datasets. **Statistical Applications in Genetics and Molecular Biology**, v. 20, n. 4, p. 456-472, 2021.

WANG, X.; YANG, Z.; LI, J. Forecasting Electricity Demand with CatBoost. **Energy Systems Analysis**, v. 12, p. 129-144, 2020.

WANG, X.; ZHANG, M. Carbon Emissions Prediction with Elastic Net. **Environmental Data Science**, v. 20, p. 112-126, 2023.

WHETSEL, K. B. Near-Infrared Spectrophotometry. **Applied Spectroscopy Reviews**, v. 2, n. 1, p. 1–68, 1968.

XU, W.; ZHANG, L.; LI, Y. Stability Analysis of Extra Trees Regressor. **Journal of Machine Learning Theory**, v. 7, p. 213-228, 2020.

YANG, Y. et al. Nitrogen fertilization weakens the linkage between soil carbon and microbial diversity: A global meta-analysis. **Global Change Biology**, v. 28, n. 21, p. 1-19, 2022.

ZHANG, J.; LIU, W.; SUN, T. Sparse data prediction using LASSO regression: New insights and methodologies. **Journal of Applied Statistics**, v. 49, n. 5, p. 234-249, 2022.

ZHANG, L.; WANG, X.; LIU, Y. Elastic Net in Genomic Studies. **Bioinformatics Journal**, v. 19, n. 2, p. 98-113, 2022.

ZHANG, T.; WANG, J. Predictive Modeling of Financial Time Series Using Extra Trees. **Journal of Financial Data Science,** v. 4, p. 129-145, 2022.

ZHANG, X.; WANG, Z. Robust Estimation with Bayesian Ridge Regression: A Review. **Journal of Computational Statistics**, v. 11, n. 2, p. 143-159, 2021.

ZHAO, C. et al. Em: ZHANG, Q. (Org). **Precision Agriculture Technology for Crop Farming.** CRC Press, 2015, p. 55-102.

ZHAO, L. et al. Distributed regularized stochastic configuration networks via the elastic net. **Neural Computing and Applications**, v. 33, n. 8, p. 3281–3297, 2020.

ZHAO, L.; WANG, K.; LIU, Y. Predicting Drug Efficacy with CatBoost in Precision Medicine. **Bioinformatics Journal**, v. 38, p. 235-249, 2022.

ZHOU, W. et al. Soil organic matter content prediction using Vis-NIRS based on different wavelength optimization algorithms and inversion models. **Journal of Soils and Sediments**, v. 23, p. 2506-2517, 2023.

APÊNDICE A – HUBER REGRESSOR

Algoritmo X. Regressor de Huber

```
INICIAR
    # Inicializar grafo direcionado para o fluxograma
    Criar um grafo G como um grafo direcionado

    # Definir os passos do algoritmo
    Definir passos:
        "Start" -> "Início: Dados de entrada (X, y)"
        "Preprocess" -> "Pré-processar os dados (verificar
outliers, normalizar, etc.)"
        "Initialize" -> "Inicializar parâmetros do modelo"
        "Define_Loss" -> "Definir a função de perda de Huber"
        "Iterate" -> "Iterar o processo de ajuste
(otimização)"
        "Update_Params" -> "Atualizar os parâmetros baseados
na minimização da perda"
        "Check_Convergence" -> "Verificar se a convergência
foi alcançada"
        "Output" -> "Saída: Modelo ajustado com parâmetros
finais"

    # Definir as conexões entre os passos
    Definir conexões:
        Conexão "Start" -> "Preprocess"
        Conexão "Preprocess" -> "Initialize"
        Conexão "Initialize" -> "Define_Loss"
        Conexão "Define_Loss" -> "Iterate"
        Conexão "Iterate" -> "Update_Params"
        Conexão "Update_Params" -> "Check_Convergence"
        Conexão "Check_Convergence" -> "Iterate"  # Loop até
convergência
        Conexão "Check_Convergence" -> "Output"  # Finaliza
após convergência
```

```
    # Adicionar passos ao grafo
    PARA cada chave, valor em passos:
        Adicionar nó ao grafo com chave e rótulo (valor)

    # Adicionar conexões ao grafo
    PARA cada conexão em conexões:
        Adicionar conexão ao grafo

    # Gerar layout para o gráfico
    Gerar posições dos nós no grafo com algoritmo de
disposição (layout)

    # Configurar visualização do fluxograma
    Criar figura de tamanho apropriado
    Desenhar nós do grafo
    Desenhar conexões com setas
    Adicionar rótulos aos nós

    # Exibir fluxograma
    Ajustar margens e exibir gráfico com título
```

FIM

APÊNDICE B – EXTRATREES REGRESSOR

Algoritmo B. ExtraTrees

Entrada:

- Dados de treinamento X (características) e y (rótulos)
- Número de árvores T
- Número de características a serem consideradas por divisão m

Saída:

- Floresta de árvores (modelo final)

Passos:

1. Inicialize a floresta como um conjunto vazio.

2. Para cada árvore t de 1 a T:
   a. Crie uma amostra aleatória do conjunto de dados de treinamento (com ou sem reposição).
   b. Construa uma árvore:
      i. Selecione aleatoriamente m características para cada nó.
      ii. Para cada característica selecionada, escolha aleatoriamente um valor como ponto de divisão dentro do intervalo dos valores observados.
      iii. Divida os dados de acordo com o melhor ponto de divisão aleatório (baseado em um critério como Gini ou entropia para classificação, ou redução do erro para regressão).
      iv. Repita o processo recursivamente até atingir um critério de parada (número mínimo de amostras em um nó ou profundidade máxima da árvore).
   c. Adicione a árvore construída à floresta.

3. Para realizar previsões:
   a. Para cada entrada de teste x:
      i. Obtenha a previsão de cada árvore na floresta.

ii. Combine as previsões:

- Para regressão: Retorne a média das previsões das árvores.

- Para classificação: Retorne a classe com a maioria dos votos.

FIM

APÊNDICE C – CATBOOST REGRESSOR

Algoritmo C. CatBoost Regressor

```
CatBoost(dataset, loss_function, num_iterations, learning_rate):
    Input:
        dataset: Conjunto de dados de entrada, incluindo variáveis categóricas e numéricas.
        loss_function: Função de perda a ser minimizada.
        num_iterations: Número de iterações (ou árvores) no _boosting_.
        learning_rate: Taxa de aprendizado para atualização dos modelos.
    Output:
        modelo_final: Modelo treinado para previsão.

    1. Pré-processamento dos dados:
        a) Identificar variáveis categóricas no dataset.
        b) Para cada variável categórica:
            i) Substituir valores categóricos por estatísticas ordenadas (_ordered statistics_):
                - Dividir os dados em subgrupos para evitar _data leakage_.
                - Calcular estatísticas de destino (_target statistics_) apenas usando os dados anteriores no subgrupo.
            ii) Converter os valores categóricos transformados para um formato numérico.

    2. Inicializar o modelo com uma predição inicial:
        modelo_0 ← Média do alvo (ou valor padrão definido pela função de perda).

    3. Para cada iteração t de 1 a num_iterations:
        a) Calcular os gradientes dos erros no conjunto de treinamento com base no modelo atual.
```

b) Treinar uma nova árvore de decisão simétrica (_symmetric tree_) usando os gradientes como alvos:

i) Construir a estrutura da árvore de forma balanceada.

ii) Dividir os dados em cada nó com base em critérios otimizados.

c) Atualizar o modelo usando a nova árvore ajustada:

modelo_t ← modelo_(t-1) + learning_rate × output_da_árvore.

4. Após todas as iterações:

a) Agregar os modelos ajustados (árvores) como um ensemble.

b) Retornar o modelo final para previsão.

Retornar: modelo_final

FIM

APÊNDICE D – ELASTIC NET

Algoritmo D. ElasticNet

```
ElasticNet(X, y, α, λ, max_iter, tol):
    Input:
        X: Matriz de variáveis independentes (dimensão n x p).
        y: Vetor de variáveis dependentes (dimensão n).
        α: Parâmetro de balanceamento entre L1 (LASSO) e L2
(Ridge) penalizações (0 ≤ α ≤ 1).
        λ: Parâmetro de regularização que controla a força da
penalização.
        max_iter: Número máximo de iterações para convergência.
        tol: Critério de tolerância para convergência.

    Output:
        β: Vetor de coeficientes ajustados (dimensão p).

    1. Inicializar os coeficientes:
        β ← vetor zero de dimensão p.

    2. Normalizar X e centralizar y:
        a) Para cada coluna de X:
            i) Subtrair a média da coluna.
            ii) Dividir pela norma Euclidiana.
        b) Subtrair a média de y.

    3. Para cada iteração t de 1 até max_iter:
        a) Salvar os coeficientes anteriores:
            β_old ← β.

        b) Atualizar cada coeficiente β_j (j de 1 a p):
            i) Calcular a predição parcial excluindo β_j:
                r_j ← y - (Xβ - X_j * β_j).
            ii) Calcular o valor não penalizado:
                z_j ← X_j^T * r_j / n.
```

iii) Atualizar β_j com a penalização ElasticNet:
β_j ← S(z_j, λ * α) / (1 + λ * (1 - α)),
onde S(z, t) é a função de _soft-thresholding_:
S(z, t) = sgn(z) * max(|z| - t, 0).

c) Verificar a convergência:
Se ||β - β_old||_2 < tol:
Encerrar as iterações.

4. Retornar os coeficientes ajustados:
Retornar β.

Fim do Algoritmo

APÊNDICE E – LASSO

Algoritmo E. LASSO

```
// LASSO: Regressão com penalização L1

// ENTRADAS
// X: matriz de variáveis independentes (n x p)
// y: vetor de variáveis dependentes (n x 1)
// lambda: parâmetro de regularização (lambda >= 0)
// tol: tolerância para critério de parada

// SAÍDA
// beta: vetor de coeficientes ajustados (p x 1)

algoritmo "LASSO"
    funcao vetor LASSO(matriz X, vetor y, real lambda, real
tol)
        inteiro n, p, t
        real dif
        vetor beta, beta_anterior, r, S
        matriz X_j

        // Inicializações
        n <- numero_linhas(X)
        p <- numero_colunas(X)
        beta <- vetor(p, 0) // Inicializa coeficientes como 0
        beta_anterior <- vetor(p, 0)
        dif <- infinito // Diferença inicial
        t <- 0

        // Loop até convergência
        enquanto (dif > tol) faca
            beta_anterior <- beta // Armazena valores
anteriores
```

```
            // Atualização por coordenadas
            para j de 1 ate p faca
                // Calcula o resíduo excluindo o coeficiente j
                r <- y - (X * beta) + (X[, j] * beta[j])
                S <- somatorio(X[, j] * r)

                // Atualiza beta[j] com penalização L1
                se (S > lambda) entao
                    beta[j] <- (S - lambda) / (somatorio(X[,
j] * X[, j]))
                senao se (S < -lambda) entao
                    beta[j] <- (S + lambda) / (somatorio(X[,
j] * X[, j]))
                senao
                    beta[j] <- 0
                fim se
            fim para

            // Calcula diferença para critério de parada
            dif <- norma(beta - beta_anterior)
            t <- t + 1
        fim enquanto

        // Retorna os coeficientes ajustados
        retorne beta
    fim da funcao
fim do algoritmo
```

APÊNDICE F – BAYESIAN RIDGE

Algoritmo F. BRR

```
// BAYESIAN RIDGE REGRESSION: Regressão Ridge Bayesiana

// ENTRADAS
// X: matriz de variáveis independentes (n x p)
// y: vetor de variáveis dependentes (n x 1)
// alpha: parâmetro de precisão da distribuição a priori dos coeficientes (prior)
// lambda: parâmetro de precisão da distribuição a priori do erro
// tol: tolerância para critério de convergência
// max_iter: número máximo de iterações

// SAÍDAS
// beta: vetor de coeficientes ajustados (p x 1)

algoritmo "BayesianRidge"
    funcao vetor BayesianRidge(matriz X, vetor y, real alpha, real lambda, real tol, inteiro max_iter)
        inteiro n, p, iteracao
        real dif
        matriz S, inv_S
        vetor beta, beta_anterior, mu

        // Inicializações
        n <- numero_linhas(X)
        p <- numero_colunas(X)
        beta <- vetor(p, 0) // Inicializa coeficientes como 0
        beta_anterior <- vetor(p, 0)
        mu <- vetor(p, 0)
        S <- matriz(p, p, 0) // Matriz de covariância inicial
        dif <- infinito
        iteracao <- 0
```

```
        // Iterações para ajuste dos coeficientes
        enquanto (dif > tol) e (iteracao < max_iter) faca
            beta_anterior <- beta // Armazena valores
anteriores

            // Calcula a matriz de covariância S = (lambda * I
+ alpha * X^T * X)^(-1)
            S <- inversa(lambda * identidade(p) + alpha *
transposta(X) * X)

            // Calcula a média mu = alpha * S * X^T * y
            mu <- alpha * S * transposta(X) * y

            // Atualiza beta (estimativa dos coeficientes)
            beta <- mu

            // Critério de parada
            dif <- norma(beta - beta_anterior)
            iteracao <- iteracao + 1
        fimenquanto

        // Retorna os coeficientes ajustados
        retorne beta
    fim da funcao
fim do algoritmo
```