

**UNIVERSIDADE ESTADUAL PAULISTA
FACULDADE DE CIÊNCIAS AGRÁRIAS E VETERINÁRIAS
CÂMPUS JABOTICABAL**

**APRENDIZADO DE MÁQUINA ESTATÍSTICO PARA ESTIMATIVA DA EMISSÃO
DE CO₂ DO SOLO EM ÁREAS AGRÍCOLAS**

Luís Felipe Trevelim

Jaboticabal - SP

**UNIVERSIDADE ESTADUAL PAULISTA
FACULDADE DE CIÊNCIAS AGRÁRIAS E VETERINÁRIAS
CÂMPUS DE JABOTICABAL**

**APRENDIZADO DE MÁQUINA ESTATÍSTICO PARA ESTIMATIVA DA EMISSÃO
DE CO₂ DO SOLO EM ÁREAS AGRÍCOLAS**

Luis Felipe Trevelim

Orientador: Prof. Dr. Alan Rodrigo Panosso

Trabalho apresentado à Faculdade de
Ciências Agrárias e Veterinárias –
UNESP, Câmpus de Jaboticabal, para
a Graduação em ENGENHARIA
AGRONÔMICA

Jaboticabal

1º Semestre/2026

FICHA CATALOGRÁFICA

T812a Trevelim, Luis Felipe

Aprendizado de máquina estatístico para estimativa da emissão de CO₂ do solo em áreas agrícolas / Luis Felipe Trevelim. -- Jaboticabal, 2026

62 p. : tabs.

Trabalho de conclusão de curso (Bacharelado - Engenharia Agrônômica) - Universidade Estadual Paulista (UNESP), Faculdade de Ciências Agrárias e Veterinárias, Jaboticabal

Orientador: Alan Rodrigo Panosso

1. Estatística. 2. Mudanças climáticas. 3. Inteligencia artificial. 4. Machine learning. 5. Ciência do solo. I. Título.

Sistema de geração automática de fichas catalográficas da Unesp. Dados fornecidos pelo autor(a).

APRENDIZADO DE MÁQUINA ESTATÍSTICO PARA ESTIMATIVA DA EMISSÃO DE CO₂ DO SOLO EM ÁREAS AGRÍCOLAS

Trabalho de Conclusão de Curso apresentado à Universidade Estadual Paulista (UNESP), Faculdade de Ciências Agrárias e Veterinárias, Jaboticabal, para obtenção do título de Bacharel em **Engenharia Agrônômica**.

Orientador: Prof. Dr. Alan Rodrigo Panosso

Coorientador (se houver):

Área de Concentração: Agronomia - Ciências do

solo Trabalho aprovado em 19/06/2026

Banca Examinadora:

Documento assinado digitalmente
 **ALAN RODRIGO PANOSSO**
Data: 22/06/2026 21:02:27-0300
Verifique em <https://validar.iti.gov.br>

Prof. Dr. Alan Rodrigo Panosso

UNESP - Faculdade de Ciências Agrárias e Veterinárias - Campus de Jaboticabal

Documento assinado digitalmente
 **GLAUCO DE SOUZA ROLIM**
Data: 22/06/2026 12:06:35-0300
Verifique em <https://validar.iti.gov.br>

Prof. Dr. Glauco de Souza Rolim

UNESP - Faculdade de Ciências Agrárias e Veterinárias - Campus de Jaboticabal

Documento assinado digitalmente
 **WITORIA DE OLIVEIRA ARAUJO**
Data: 22/06/2026 12:15:05-0300
Verifique em <https://validar.iti.gov.br>

Me. Witória de Oliveira Araújo

UNESP - Faculdade de Ciências Agrárias e Veterinárias - Campus de Jaboticabal

Documento assinado digitalmente
 **GLAUCO DE SOUZA ROLIM**
Data: 22/06/2026 12:08:29-0300
Verifique em <https://validar.iti.gov.br>

Prof. Dr. Glauco de Souza Rolim

Chefe do Departamento

DADOS CURRICULARES DO AUTOR

Luis Felipe Trevelim – Nascido em São José do Rio Preto, São Paulo, no dia 1 de agosto de 2003, filho de Claudemir Trevelim Junior e Andressa Fabiana Miola. Cursou Engenharia Agrônômica, pela Faculdade de Ciências Agrárias e Veterinárias da Universidade Estadual Paulista (FCAV-UNESP), Campus de Jaboticabal-SP, entre janeiro de 2022 e dezembro de 2026. Durante este período estagiou no Grupo de Estudos em Estatística Computacional Aplicada (GEECA), no qual adquiriu conhecimento nas áreas de estatística, processamento e análise de dados. No ano de 2025 fez parte do Grupo de Integração Empresa Universidade (GIEU) da UNESP/FCAV.

DEDICATÓRIA

Dedico este trabalho aos meus pais, Claudemir Trevelim Junior e Andressa Fabiana Miola, pelo amor incondicional, pelo apoio em cada etapa da minha formação e por nunca medirem esforços para que eu pudesse chegar até aqui.

À minha irmã Lara Ândrea Trevelim, pela companhia, pelo carinho e por tornar os dias mais leves.

À minha namorada Yasmin de Carvalho Levati, pela paciência, pelo incentivo e por estar ao meu lado nos momentos mais difíceis desta jornada.

*"A educação é a arma mais poderosa que você pode usar para mudar o mundo." —
Nelson Mandela"*

AGRADECIMENTOS

Agradeço primeiramente aos meus pais, Claudemir Trevelim Junior e Andressa Fabiana Miola, e à minha irmã Lara Ândrea Trevelim, pelo suporte incondicional ao longo de toda a minha trajetória acadêmica. À minha namorada Yasmin de Carvalho Levati, pelo companheirismo, pela paciência e pelo apoio nos momentos mais desafiadores desta etapa.

Ao meu orientador, Prof. Dr. Alan Rodrigo Panosso, pela orientação dedicada, pelos ensinamentos, pela confiança depositada no desenvolvimento deste trabalho e por sempre indicar os melhores caminhos com competência e generosidade.

À Faculdade de Ciências Agrárias e Veterinárias da Universidade Estadual Paulista "Júlio de Mesquita Filho" (FCAV – UNESP), Campus de Jaboticabal, pela estrutura, pelo ambiente acadêmico e pelas oportunidades proporcionadas ao longo da minha formação.

Ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) e à Fundação de Amparo à Pesquisa do Estado de São Paulo – FAPESP, pelo auxílio financeiro para desenvolvimento desta pesquisa - 2024/12296-6

E a Deus, por me conceder saúde, força e determinação para superar os desafios.

SUMÁRIO

RESUMO	iii
ABSTRACT	iv
1 INTRODUÇÃO	6
2 REVISÃO DE LITERATURA	7
2.1 Mudanças climáticas e emissões de CO ₂	7
2.2 Emissão de CO ₂ do solo (FCO ₂)	7
2.3 Variáveis ambientais e modelagem da FCO ₂	8
2.4 Uso de aprendizado de máquina em estudos ambientais	8
2.5 Importância da vegetação na dinâmica do carbono	9
3 MATERIAL E MÉTODOS	10
3.1 Banco de dados legados	10
3.2 Medição das emissões de CO ₂ , temperatura e umidade do solo	10
3.3 Determinação dos atributos químicos e físicos do solo	10
3.4 Aquisição de dados da estação agrometeorológica	11
3.5 Aquisição de produtos de mapeamento digital	11
3.6 Processamento e análise dos dados	12
3.7 Aprendizado de máquina estatístico	Error! Bookmark not defined.
3.8 Análise de Explicabilidade dos Modelos (SHAP)	17
4 RESULTADOS E DISCUSSÕES	18
4.1 Configuração e regularização dos hiperparâmetros	18
4.2 Desempenho do Modelo	31
4.3 Importância das variáveis para os modelos	34
4.4 Análise de Explicabilidade dos Modelos (SHAP)	37
5 CONSIDERAÇÕES FINAIS	43
6 REFERÊNCIAS	45

RESUMO

A concentração de gases de efeito estufa (GEE) na atmosfera, como o dióxido de carbono (CO_2), aumentou consideravelmente devido a fontes antropogênicas. No Brasil, atividades agrícolas e florestais contribuem substancialmente para as emissões de CO_2 , principalmente devido ao desmatamento e à conversão de florestas nativas. Estudos anteriores demonstraram que FCO_2 pode ser modelada com alta precisão usando uma grande quantidade de variáveis ambientais. No entanto, a conversão a longo prazo de florestas nativas para agroecossistemas ainda é pouco compreendida, especialmente no contexto brasileiro. Assim, a hipótese central deste estudo é que modelos de aprendizado de máquina baseados em *ensembles* de árvores, como *Random Forest* e *Extreme Gradient Boosting*, apresentam desempenho preditivo superior aos demais algoritmos avaliados para a estimativa da FCO_2 em áreas agrícolas do Cerrado. Este projeto investigou a emissão de CO_2 do solo (FCO_2) em áreas agrícolas do bioma Cerrado, utilizando técnicas de aprendizado de máquina estatístico para modelar FCO_2 com base em demais variáveis associadas. As técnicas utilizadas foram: redes neurais artificiais (*RNA*), *Support Vector Machine* (*SVM*), *Random Forest* (*RF*), *Decision Tree* (*DT*) e *Extreme Gradient Boosting* (*XGB*). De maneira geral 70-80% das observações foram utilizadas para aprendizagem (processo de treinamento) dos modelos e 30-20% para validação. A acurácia dos modelos foi determinada por meio da correlação de Pearson (r), coeficiente de determinação (R^2), erro quadrático médio (*RMSE*) e menor erro percentual absoluto médio (*MAPE*). Os modelos baseados em árvores *ensemble* apresentaram os melhores desempenhos. O *Random Forest* obteve o maior poder explicativo ($R^2 = 0,69$) e o menor erro ($\text{RMSE} = 1,18 \mu\text{mol m}^{-2} \text{s}^{-1}$), seguido de perto pelo *XGB* ($R^2 = 0,66$; $\text{RMSE} = 1,21 \mu\text{mol m}^{-2} \text{s}^{-1}$). Os modelos *SVM* apresentaram desempenho intermediário ($R^2 = 0,64$), enquanto *RNA* e *DT* mostraram menor capacidade preditiva ($R^2 = 0,61$). A análise de importância das variáveis, complementada pela análise de explicabilidade *SHAP* (*SHapley Additive exPlanations*), revelou que indicadores da atividade da vegetação, como a fluorescência induzida pelo sol (*SIF*), além de variáveis climáticas como temperatura e umidade, estiveram entre os principais preditores em todos os modelos. A recorrência dessas variáveis reforça o papel conjunto da atividade biológica, das condições térmicas e hídricas do solo na regulação da respiração do solo. Essa abordagem contribuiu para a compreensão da

dinâmica do fluxo de CO₂ em solos brasileiros e forneceu bases para a construção de padrões espaçotemporais mais precisos, essenciais para o desenvolvimento de estratégias eficazes de mitigação das mudanças climáticas. **Palavras-chaves:** respiração do solo, inteligência artificial, mudanças climáticas, aprendizado de máquina.

ABSTRACT

The concentration of greenhouse gases (GHG) in the atmosphere, such as carbon dioxide (CO₂), has increased considerably due to anthropogenic sources. In Brazil, agricultural and forestry activities contribute substantially to CO₂ emissions, mainly due to deforestation and the conversion of native forests. Previous studies have demonstrated that FCO₂ can be modeled with high accuracy using a large number of environmental variables. However, the long-term conversion of native forests to agroecosystems remains poorly understood, especially in the Brazilian context. Thus, the central hypothesis of this study is that ensemble tree-based machine learning models, such as Random Forest and Extreme Gradient Boosting, exhibit superior predictive performance compared to the other evaluated algorithms for estimating FCO₂ in agricultural areas of the Cerrado. This study investigated soil CO₂ emission (FCO₂) in agricultural areas of the Cerrado biome using statistical machine learning techniques to model FCO₂ based on associated environmental variables. The techniques applied were: artificial neural networks (ANN), Support Vector Machine (SVM), Random Forest (RF), Decision Tree (DT), and Extreme Gradient Boosting (XGB). In general, 70–80% of the observations were used for model training and 20–30% for validation. Model accuracy was assessed using Pearson's correlation coefficient (r), coefficient of determination (R^2), root mean square error (RMSE), and mean absolute percentage error (MAPE). Ensemble tree-based models achieved the best overall performance. Random Forest yielded the highest explanatory power ($R^2 = 0.69$) and the lowest error (RMSE = $1.18 \mu\text{mol m}^{-2} \text{s}^{-1}$), followed closely by XGB ($R^2 = 0.66$; RMSE = $1.21 \mu\text{mol m}^{-2} \text{s}^{-1}$). SVM models showed intermediate performance ($R^2 = 0.64$), while ANN and DT exhibited lower predictive capacity ($R^2 = 0.61$). Variable importance analysis, complemented by SHAP (SHapley Additive exPlanations) explainability analysis, revealed that vegetation activity indicators, such as solar-induced fluorescence (SIF), along with climatic variables such as temperature and soil

moisture, were among the main predictors across all models. The recurrence of these variables highlights the combined role of biological activity and soil thermal and hydric conditions in regulating soil respiration. This approach contributed to understanding the dynamics of CO₂ flux in Brazilian soils and provided a foundation for building more accurate spatiotemporal patterns, essential for the development of effective climate change mitigation strategies.

Keywords: soil respiration, artificial intelligence, climate change, machine learning.

1 INTRODUÇÃO

As mudanças climáticas vêm se consolidando como um dos principais desafios ambientais da atualidade, especialmente em função do aumento da concentração de gases de efeito estufa na atmosfera. Dentre esses gases, o dióxido de carbono (CO_2) desempenha papel central, por estar diretamente associado às atividades humanas, como o desmatamento e o uso intensivo do solo, o que torna cada vez mais importante compreender como diferentes sistemas contribuem para o balanço de carbono. No Brasil, essa discussão ganha maior relevância devido à forte presença do setor agropecuário na economia. Regiões como o Cerrado, que ocupam posição estratégica na produção agrícola nacional, têm passado por intensas transformações no uso da terra ao longo das últimas décadas. Essas mudanças afetam diretamente o funcionamento do solo e os processos envolvidos na troca de carbono com a atmosfera. A emissão de CO_2 do solo, também conhecida como respiração do solo, é um dos principais processos responsáveis pela liberação de carbono para a atmosfera. Esse processo resulta da interação entre a atividade microbiana e a respiração das raízes das plantas, sendo influenciado por fatores como temperatura, umidade e atributos físico-químicos do solo. Por se tratar de um processo dinâmico, sua variação ao longo do tempo e do espaço torna sua compreensão um desafio. Apesar dos avanços na área, ainda existem limitações na capacidade de explicar e prever a emissão de CO_2 do solo, especialmente em ambientes tropicais, onde as condições ambientais são mais variáveis. A presença de múltiplos fatores atuando simultaneamente dificulta a identificação dos principais controladores desse processo, tornando necessária a utilização de abordagens que consigam lidar com essa complexidade. Nesse contexto, o uso de técnicas de aprendizado de máquina tem se mostrado uma alternativa promissora, permitindo integrar diferentes variáveis ambientais e identificar padrões complexos, contribuindo para uma melhor compreensão dos processos que regulam a dinâmica do carbono no solo. Diante disso, este trabalho tem como objetivo analisar a emissão de CO_2 do solo em áreas agrícolas do Cerrado, considerando a influência de variáveis ambientais e utilizando modelos de aprendizado de máquina para descrever e prever esse processo.

2 REVISÃO DE LITERATURA

2.1 Mudanças climáticas e emissões de CO₂

O aumento da concentração de gases de efeito estufa na atmosfera, especialmente o dióxido de carbono (CO₂), está relacionado às atividades antrópicas, com destaque para os setores de Agricultura, Silvicultura e Uso da Terra (Paustian et al., 2016; IPCC, 2013). No Brasil, essas emissões estão fortemente associadas ao desmatamento e à conversão de ecossistemas naturais em áreas agrícolas, principalmente no bioma Cerrado (Cerri et al., 2007). O setor agropecuário brasileiro é responsável por parcela significativa das emissões nacionais de GEE. Segundo o Sistema de Estimativas de Emissões de Gases de Efeito Estufa (SEEG, 2018), as atividades relacionadas à agropecuária e ao uso da terra respondem pela maior parte das emissões do país. Nesse contexto, compreender os mecanismos de troca de carbono entre o solo e a atmosfera torna-se fundamental para a elaboração de estratégias de mitigação mais eficazes (Paustian et al., 2016). A concentração atmosférica de CO₂ vem sendo monitorada de forma contínua por satélites como o *Orbiting Carbon Observatory-2* (OCO-2) da NASA, que mede indiretamente a concentração da coluna de CO₂ (XCO₂) por meio da intensidade da radiação solar refletida (Crisp et al., 2012; Massie et al., 2021). Esses dados orbitais têm ampliado a capacidade de monitoramento em escala regional e global, complementando as medições realizadas em campo (da Costa et al., 2021; Moraes Filho et al., 2021).

2.2 Emissão de CO₂ do solo (FCO₂)

A emissão de CO₂ do solo (FCO₂), também conhecida como respiração do solo, resulta da interação entre a atividade microbiana e a respiração radicular (Ussiri e Lal, 2009). Esse processo é controlado por variáveis ambientais, como temperatura, umidade e atributos físico-químicos do solo, que influenciam tanto a produção quanto o transporte de gases (Silva et al., 2019; Xavier et al., 2020). A FCO₂ é considerada a segunda maior fonte de CO₂ para a atmosfera, ficando atrás apenas dos oceanos (Panosso, 2023). Pequenas alterações nas taxas de emissão podem ser suficientes para que um ecossistema mude de sumidouro para fonte de carbono, o que evidencia a importância do monitoramento preciso desse processo. A respiração do solo é composta por duas frações principais: a respiração autotrófica, associada às raízes das plantas e fungos micorrízicos, e a respiração heterotrófica, decorrente da decomposição da matéria orgânica por microrganismos do solo (Ussiri e Lal, 2009).

Na escala temporal, a FCO_2 é controlada principalmente pela umidade e temperatura do solo (Moitinho et al., 2015; Moitinho et al., 2018). Na escala espacial, atributos físico-químicos como porosidade, densidade do solo, pH e teores de carbono orgânico exercem papel importante na variabilidade observada.

2.3 Variáveis ambientais e modelagem da FCO_2

Estudos indicam que a FCO_2 pode ser modelada a partir de múltiplas variáveis ambientais, incluindo propriedades do solo e fatores climáticos (Hamrani, Akbarzadeh e Madramootoo, 2020; Canteral et al., 2023). Abordagens que integram diferentes variáveis ambientais, como solo, clima, relevo e biota, têm se mostrado mais eficientes para explicar a dinâmica do carbono (McBratney, Santos e Minasny, 2003; Grunwald, 2022). A temperatura do solo é reconhecida como um dos principais controladores da FCO_2 ao longo do tempo, por regular diretamente a atividade enzimática e microbiana no solo (Moitinho et al., 2015; Vicentini et al., 2023). A umidade do solo, por sua vez, atua sobre a difusão de gases no perfil e sobre as condições aeróbicas necessárias para a decomposição da matéria orgânica. Estudos conduzidos em áreas de cana-de-açúcar no Brasil demonstraram que essas duas variáveis explicam grande parte da variabilidade temporal da FCO_2 (Tavares et al., 2019; Farhate et al., 2018). Além das variáveis de solo, dados provenientes de estações agrometeorológicas, como temperatura do ar, radiação fotossinteticamente ativa (PAR), evapotranspiração e velocidade do vento, têm sido incorporados em modelos preditivos de FCO_2 , ampliando a capacidade explicativa dos algoritmos (Canteral et al., 2023; Panosso, 2023).

2.4 Uso de aprendizado de máquina em estudos ambientais

O uso de técnicas de aprendizado de máquina tem se destacado por sua capacidade de lidar com dados complexos e identificar relações não lineares entre variáveis (Reichstein et al., 2019). Essas abordagens permitem integrar múltiplos fatores ambientais, melhorando a capacidade preditiva dos modelos em sistemas dinâmicos como o solo. Na última década, observou-se aumento expressivo da aplicação de algoritmos de aprendizado de máquina para modelagem da FCO_2 em diferentes condições de solo e manejo (Freitas et al., 2018; Canteral et al., 2023). Algoritmos como Redes Neurais Artificiais (RNA), *Random Forest* (RF), *Support Vector Machine* (SVM), Árvores de Decisão (DT) e *Extreme Gradient Boosting* (XGB)

têm sido avaliados em diferentes contextos, apresentando desempenhos variáveis conforme o conjunto de variáveis e a escala de análise adotada (Panosso, 2023; Harsányi et al., 2024). Entre os algoritmos avaliados na literatura, o *RF* e o *XGB* destacam-se por sua robustez e capacidade de lidar com grandes volumes de dados e variáveis de diferentes naturezas. O *RF* combina múltiplas árvores de decisão geradas a partir de amostras aleatórias dos dados, reduzindo o risco de *overfitting* e aumentando a estabilidade das predições (Breiman et al., 2022). O *XGB*, por sua vez, utiliza o conceito de *boosting*, construindo árvores sequencialmente de forma a corrigir os erros das anteriores, o que resulta em modelos com alta acurácia preditiva (James et al., 2013; Kroese et al., 2019).

2.5 Importância da vegetação na dinâmica do carbono

A inclusão de variáveis relacionadas à vegetação tem contribuído para avanços na compreensão dos fluxos de carbono. Indicadores como a fluorescência induzida pelo sol (SIF) apresentam forte relação com a dinâmica do CO_2 , relacionando a conexão entre a atividade fotossintética e a emissão de CO_2 do solo (Costa et al., 2022). A SIF é um sinal eletromagnético emitido pela vegetação durante o processo de fotossíntese, captado por satélites como o OCO-2, e apresenta alta correlação com a produtividade primária bruta (GPP) dos ecossistemas (da Costa et al., 2023; Mohammed et al., 2019). Sua inclusão como variável preditora em modelos de FCO_2 tem permitido capturar a influência da atividade vegetal sobre os processos de respiração do solo, especialmente por meio do aporte de carbono recente via exsudatos radiculares e deposição de material orgânico (da Costa et al., 2021). Além da SIF, a concentração atmosférica de CO_2 (XCO_2) obtida por sensoriamento remoto tem sido utilizada como variável auxiliar na modelagem da FCO_2 , uma vez que reflete as condições de troca gasosa entre o solo, a vegetação e a atmosfera em escala regional (Moraes Filho et al., 2021; da Costa et al., 2022).

3 MATERIAL E MÉTODOS

3.1 Banco de dados legados

Este trabalho fundamenta-se em uma compilação de dados históricos provenientes de ensaios realizados no Brasil nos últimos 12 anos. A base de dados utilizada deriva do repositório estruturado por Panosso (2023), que engloba experimentos conduzidos em diversas culturas e condições de manejo. Foram selecionados estudos que detalham não apenas o fluxo de CO₂, mas também variáveis auxiliares como temperatura e umidade, totalizando uma série de observações que refletem as transformações no uso da terra em regiões como São Paulo e Mato Grosso do Sul (Tavares et al., 2015; Campos et al., 2016; Mantovanelli et al., 2016; Oliveira, 2016; Bicalho et al., 2017; de Figueiredo et al., 2017; Almeida et al., 2018; da Cunha et al., 2018; Farhate et al., 2018; Freitas et al., 2018; Tavares et al., 2018; Pinheiro da Silva et al., 2019; Vicentini et al., 2019; Tavanti et al., 2020; Moitinho et al., 2021a; Moitinho et al., 2021b; Moitinho et al., 2021c; Costa et al., 2022; da Costa et al., 2022; Rossi et al., 2022; da Costa et al., 2023; De Lucena et al., 2023; de Souza Maria et al., 2023).

3.2 Medição das emissões de CO₂, temperatura e umidade do solo

As avaliações da emissão de CO₂ do solo (FCO₂), temperatura do solo (Ts) e umidade do solo (Us) foram realizadas entre 7 e 12 h. Foram estabelecidos 25 pontos de amostragem para a FCO₂, Ts e Us, por meio de colares de cloreto de polivinila (PVC) com diâmetro de 0,10 m e altura de 0,085 m, instalados previamente nas áreas, permanecendo fixos até o término do experimento.

A FCO₂ foi registrada por 1,5 minutos em cada ponto amostral, por meio de um sistema de fluxo de solo (LI-8100; LI-COR Bioscience, NE, EUA), que monitora as mudanças na concentração de CO₂ dentro da câmara para solos por espectroscopia de absorção óptica na região do infravermelho. A Ts foi determinada para todos os pontos amostrais usando um termômetro digital e a Us foi medida por meio de um equipamento de reflectometria no domínio do tempo (TDR) (*HydrosenseTM Campbell Scientific*, Austrália).

3.3 Determinação dos atributos químicos e físicos do solo

Ao final das avaliações de FCO₂, Ts e Us, foram coletadas amostras de solo deformadas e não deformadas nas profundidades de 0-10, 10-20 e 20-40 cm nos locais próximos aos colares de PVC para determinação dos atributos químicos e

físicos do solo. As análises químicas incluíram os seguintes atributos do solo: teores de cálcio (Ca^{2+}), magnésio (Mg^{2+}), potássio (K^+), alumínio trocável (Al^{3+}), acidez potencial ($\text{H} + \text{Al}$), fósforo disponível (P) e matéria orgânica do solo (MOS), sendo determinados de acordo com Raji et al. (2001). As análises físicas do solo compreenderam os seguintes atributos: distribuição granulométrica (areia, silte e argila), densidade do solo (D_s), volume total de poros, microporosidade e macroporosidade (EMBRAPA, 1997).

3.4 Aquisição de dados da estação agrometeorológica

Em adição, as seguintes variáveis climatológicas foram incorporadas à análise: velocidades do vento (V_{elmax} e V_{elmin} em m s^{-1}); evapotranspiração por Penman_Monteith (E_{to} em mm dia^{-1}); radiação fotossinteticamente ativa (PAR em $\mu\text{mol m}^{-2} \text{s}^{-1}$); radiação global (Rad em $\text{MJ m}^{-2} \text{dia}^{-1}$); pressão atmosférica (P_{kpa} em kpa); temperaturas do ar média, máxima e mínima (T_{med} , T_{max} e T_{min} , respectivamente, em $^{\circ}\text{C}$), precipitação atmosférica (Chuva em mm), umidade relativas média, máxima e mínima (U_{med} , U_{max} e U_{min} , respectivamente, em %), insolação (Inso em h dia^{-1}). A radiação (PAR) é a radiação incidente na faixa de ondas de 400 a 700 nm, que pode ser absorvida pelo sistema fotossintético das plantas. Seu valor é responsável por aproximadamente 50% da radiação solar. As variáveis foram utilizadas a partir de estações agrometeorológicas localizadas em municípios próximos às regiões de determinação da emissão de CO_2 do solo.

3.5 Aquisição de produtos de mapeamento digital

Gases de efeito estufa (GEE), clima e dados de vegetação foram coletados de diferentes satélites ao longo de uma série temporal de 2015 a 2023 e foram agregados em escala mensal. O principal produto do *Orbiting Carbon Observatory-2* (OCO-2) consiste em estimativas georreferenciadas da concentração média de CO_2 atmosférico (XCO_2). Além disso, a fluorescência de clorofila induzida pelo sol (SIF) foi obtida devido à sobreposição que ocorre nas faixas de comprimento de onda do SIF com a faixa de absorção de O_2 (680–850 nm) (Crisp et al., 2012; O'Dell et al., 2012; Mohammed et al., 2019). Dados deste satélite já foram validados por Crisp et al. (2012) e, de acordo com O'Dell et al. (2012), esse satélite fornece cerca de 65.000 observações de qualidade por dia em todo o mundo. Foi utilizada a versão 9 do OCO-2 com uma correção de viés e foram considerados somente as medições de melhor

qualidade, indicada pela flag = 0, o que significa que não ocorre cobertura de nuvens, (Nikitenko et al., 2020; Massie et al., 2021). Em relação ao SIF, levamos em conta apenas o SIF em 757 nm, isso se deve a estudos anteriores que exploraram a relação no Estado de São Paulo (da Costa et al., 2021; Moraes Filho et al., 2021), além do fato de que esse comprimento de onda é o mais próximo do pico de vermelho distante (~740 nm) em todo o sinal de SIF (Mohammed et al., 2019).

Os dados do sensor MODIS foram extraídos do *Application for Extracting and Exploring Analysis Ready Samples* (AppEEARS). Esta aplicação permite que os usuários obtenham subconjuntos de grandes bancos de dados usando parâmetros espaciais e temporais. Dois tipos de solicitações de amostras estão disponíveis: amostras de pontos, inserindo coordenadas geográficas, e amostras de áreas utilizando polígonos vetoriais. As solicitações de amostras enviadas ao AppEEARS fornecem aos usuários não apenas os dados, mas também os valores de dados de qualidade associados. Visualizações interativas com estatísticas resumidas serão fornecidas para cada amostra dentro da aplicação, o que permite visualização e interação entre as amostras antes da aquisição específica (AppEEARS Team, 2020). Já, os dados do Nasa Power (<https://power.larc.nasa.gov>) consistem em precipitação (mm), irradiância solar de onda curta na superfície ($\text{MJ m}^{-2} \text{ dia}^{-1}$), temperatura média do ar ($^{\circ}\text{C}$) e umidade relativa a 2 m (%). Esta plataforma é composta pelo projeto da NASA intitulado: *Worldwide Energy Resource Forecast* (POWER) e foi iniciado para aprimorar o conjunto de dados de energia renovável atual e criar conjuntos de dados a partir de novos sistemas de satélites (Stackhouse et al., 2015).

3.6 Processamento e análise dos dados

Todas as análises estatísticas e as etapas de seleção de variáveis, pré-processamento, aprendizado de máquina estatístico, validação e avaliação do desempenho dos modelos, foram realizadas a partir de implementações em linguagem R (R Development Core Team, 2023) utilizando o ambiente de desenvolvimento RStudio m (Posit team, 2023). Foram utilizados os meta-pacotes *tidyverse* (Wickham et al., 2019) e *tidymodels* (Kuhn e Wickham, 2020). Além disso, aplicou-se estatística multivariada para melhor compreender as relações entre as variáveis de estudo e utilizar os algoritmos de Redes Neurais Artificiais (RNA), *Support Vector Machine* (SVM), *Random Forest* (RF), *Decision Tree* (DT) e *Extreme Gradient Boosting* (XGB) para estimar a emissão de CO_2 do solo a partir dos atributos físico-

químicos, utilizando modelos de aprendizado de máquina derivados por Canteral et al. (2023).

As etapas de pré-processamento e mineração de dados, que foram utilizados para o processo de aprendizado de máquina estatístico, foram derivadas do clássico protocolo CRISP-DM (*Cross Industry Standard Process fo Data Mining*) adaptado de Chapman et al. (2000). Inicialmente, na fase de Compreensão, os objetivos foram delineados focando na predição da FCO₂ em sistemas agrícolas tropicais. No Entendimento dos dados, priorizou-se a integração de informações multitemporais e multiespaciais vindas de sensores e estações climáticas. A Preparação envolveu o tratamento de inconsistências e a normalização de variáveis, enquanto a Modelagem aplicou os algoritmos de aprendizado supervisionado para capturar relações não lineares. Por fim, a Avaliação utilizou métricas de erro para validar a capacidade de generalização dos modelos, culminando na Implantação, que visa a disseminação dos resultados para a comunidade científica. Para avaliação dos modelos, foi primeiramente necessário a separação do banco de dados em duas partes, sendo uma para o processo de aprendizado e outra para o teste dos modelos gerados a qual foi feita pelo processo de validação cruzada. De maneira geral 70-80% das observações foram utilizadas para a aprendizagem estatística dos modelos (processo de treinamento) e 30-20% para validação dos modelos (processo de teste). Implantação: proposição de aplicação e distribuição do conhecimento gerado durante o processo. Os resultados foram organizados para divulgação de forma que este possa ser utilizado pela comunidade principalmente na forma de trabalhos e comunicações científicas. A seguir serão apresentados os principais algoritmos de aprendizado de máquina utilizados no presente projeto.

3.7 Modelos de aprendizagem de máquina

Redes Neurais Artificiais

Neste estudo, foram avaliados modelos de RNA fundamentados em três arquiteturas distintas: a rede neural de regressão generalizada (*GRNN*), os *multilayer perceptrons* (*MLP*) e as redes de função de base radial (*RBF*) (Specht, 1991; Haykin, 2001; Boateng; Pillay; Davis, 2019). A rede *GRNN* caracteriza-se como uma generalização das redes *RBF* e probabilísticas (*PNN*), apresentando uma estrutura invariável composta por uma camada de entrada, duas intermediárias e uma de saída, operando sob treinamento supervisionado com critérios de erro pré-definidos. As

redes do tipo *MLP* integram múltiplos *perceptrons* e utilizam o algoritmo de retropropagação (*backpropagation*) para realizar o ajuste fino de pesos e bias em relação ao erro, que pode ser medido de diversas maneiras (MSE, por exemplo). Estruturalmente, as *MLPs* contam com uma camada de entrada para recepção de sinais, uma camada de saída para a tomada de decisão e camadas ocultas que atuam como o motor computacional do sistema. Já as redes *RBFs* apresentam uma estrutura simples em termos de direção do fluxo de informações, possuindo uma camada de entrada, uma camada oculta e uma camada de saída.

Support Vector Machine

O *Support Vector Machine* (SVM) atua como um método de aprendizado supervisionado para problemas de classificação e regressão (Eq. 7), sendo amplamente reconhecido por sua natureza não linear (Harsányi et al., 2024). Engloba funções ou classificadores que implementam modificações espaciais nos dados, levando-os a um plano onde a classificação se torna mais fácil (hiperplanos). Para certo dado de treinamento representado por $(x_1, y_1); \dots; (x_N, y_N)$, onde x é um vetor de entrada e y é um vetor de saída correspondente, a função de regressão do SVM pode ser expressa como:

$$f(x) = \omega \cdot \phi(x) + b, \quad (7)$$

onde ω é um vetor de peso, b é um viés e ϕ apresenta uma função de transferência não linear.

O objetivo de ϕ é mapear de maneira não-linear os vetores de entrada para um espaço de característica de alta dimensão (Harsányi et al., 2024). A precisão do modelo SVM com Função Base Radial (RBF), por exemplo, depende das seleções dos parâmetros C (custo), e regularização positivo que corresponde aos parâmetros do peso no problema de otimização para minimizar o erro empírico; a variável de penalização do treinamento ε (épsilon) e do parâmetro do kernel γ (gamma). Foram inicialmente selecionadas a regressão SVM tipo 1 e função de Base Radial (RBF) e SVM tipo 2 de base Linear (LBF) de kernel. Os hiperparâmetros C , γ e ε foram ajustados para o modelo SVM tipo 1 e $(C$ e $\varepsilon)$ para SVM tipo 2. Assim, muitas combinações desses parâmetros foram testadas a partir de algoritmo de busca aleatória. O algoritmo de validação cruzada foi utilizado para obter melhores estimativas de treinamento de parâmetros do modelo SVM escolhido.

Decision Tree (DT)

As árvores de decisão constituem uma abordagem estatística de aprendizado de máquina amplamente empregada em problemas de classificação e regressão. Neste estudo, a implementação foi realizada por meio do pacote *rpart* do software R (Therneau; Atkinson; Ripley, 2022). De forma geral, o método constrói uma estrutura hierárquica em formato de árvore, na qual o conjunto de dados é particionado em subconjuntos com base nas variáveis preditoras. Esse procedimento permite capturar relações não lineares presentes nos dados, além de possibilitar o uso de variáveis categóricas.

O processo de construção da árvore ocorre de maneira recursiva, sendo que, em cada etapa, o algoritmo seleciona a variável que proporciona a maior redução do erro de regressão para realizar a divisão dos dados (James et al., 2013; Kroese et al., 2019). Entre as principais vantagens dessa técnica destacam-se a facilidade de interpretação dos resultados e a relativa robustez à presença de valores extremos, uma vez que esses modelos tendem a ser menos sensíveis a *outliers*. Durante o processo de ajuste, foram otimizados os seguintes hiperparâmetros: o custo de complexidade da árvore (*cost_complexity*), a profundidade máxima da árvore (*tree_depth*) e o número mínimo de observações por nó terminal (*min_n*). Ressalta-se que as árvores de decisão constituíram a base para a construção dos algoritmos do tipo *ensemble* utilizados neste trabalho, nos quais múltiplas árvores foram combinadas com o objetivo de aprimorar o desempenho preditivo dos modelos avaliados.

Random Forest

São modelos estruturados em árvore utilizados tanto para classificação, quanto para regressão, sendo compostos por uma combinação de árvores preditoras (floresta) geradas a partir de um vetor aleatório, utilizando uma técnica chamada *Bagging (Bootstrap Aggregating)*, amostrado de forma independente e com a mesma distribuição para todas as árvores na floresta (James et al., 2013; Kroese et al., 2019). De forma genérica, o *Bagging* baseia-se na construção de múltiplas árvores de decisão independentes, onde cada árvore é treinada em um subconjunto aleatório e com substituição do conjunto de dados de treinamento original. Na construção de uma árvore de regressão o modelo se baseia na divisão das features, conforme suas semelhanças e dissimilaridades e o seu resultado é formado pela média dos

resultados de todas as árvores. Em outras palavras o algoritmo treina múltiplas árvores de decisão em subconjuntos aleatórios dos dados e combina suas previsões. É uma técnica não paramétrica de aprendizagem estatística. A implementação foi realizada a partir do pacote em R *randomForest* (Breiman et al., 2022). No seu processo de aprendizagem, três hiperparâmetros foram otimizados: o número de árvores no *Random Forest* (*tree*); número mínimo de observações por folha (*min_n*) e o número de features consideradas em cada nó de divisão (*mtry*).

Extreme Gradient Boosting (XGB)

O *Boosting* baseia-se no conceito de melhorar algoritmos de classificação/regressão fracos, transformando-os em algoritmos mais fortes (James et al., 2013; Kroese et al., 2019). Assim como no algoritmo *Random Forest*, ele distribui pesos para as árvores individuais e, em seguida, gera um consenso combinando suas saídas. Nessa abordagem, um classificador recebe um peso maior se seu antecessor apresentar um desempenho pior, e esses pesos são ajustados durante o processo de aprendizado estatístico. O método funciona aplicando-se sequencialmente os classificadores às versões re-ponderadas do conjunto de treinamento, atribuindo maior peso às observações classificadas incorretamente no passo imediatamente anterior e menor peso às classificadas corretamente. Isso resulta em uma sequência de árvores que utilizam informações das árvores anteriores, permitindo ao algoritmo aprender com seus erros. A implementação foi realizada a partir do pacote em R *XGBoost* (Chen et al., 2023). No seu processo de aprendizagem, sete hiperparâmetros foram otimizados, divididos em cinco etapas de ajuste: etapa 1: taxa de aprendizado (*learn_rate*) e o número de árvores a serem treinadas (*trees*); etapa 2: profundidade de árvore (*tree_depth*) e número mínimo de observações por folhas (*min_n*); etapa 3: somente a redução de perda mínima para divisão (*loss_reduction*); etapa 4: o número de variáveis consideradas em cada nó de divisão (*mtry*) e o tamanho da amostra aleatória usada em cada iteração (*sample_size*); etapa 5: os hiperparâmetros *trees* e *learn_rate* foram novamente otimizados.

Para cada um dos algoritmos anteriormente descritos nesta aplicação, uma lista de valores candidatos para seus hiperparâmetros foi definida. Em seguida, utilizando o procedimento de validação cruzada *k*-fold com $k = 5$, foi realizada a análise de desempenho preditivo para cada modelo, avaliando a raiz do erro quadrático médio (RMSE – Eq. 4). Tal procedimento permitiu selecionar o modelo com

o melhor desempenho em cada iteração, considerando os dados espaciais e temporais em diferentes configurações de variáveis, passos referentes ao aprendizado estatístico e predição de FCO₂. Posteriormente, o modelo selecionado foi aplicado aos dados de teste para avaliar seu desempenho na previsão de observações futuras, utilizando novamente o RMSE como métrica de avaliação. Os diferentes modelos foram comparados pelos seguintes indicadores estatísticos: coeficiente de correlação (r – Eq. 1), coeficiente de determinação (R^2 – Eq. 2), erro quadrático médio (MSE – Eq. 3), raiz do erro quadrático médio (RMSE – Eq. 4), erro médio absoluto (MAE – Eq. 5), erro absoluto médio percentual (MAPE – Eq. 6).

$$r = \frac{\sum_{i=1}^n \sigma_{ht_i} \times \sigma_{hp_i} - \bar{\sigma}_{ht} \times \bar{\sigma}_{hp}}{(n-1) S_{\sigma_{ht}} S_{\sigma_{hp}}}, \quad (1)$$

$$R^2 = 1 - \frac{\sum_i (\sigma_{hp_i} - \sigma_{ht_i})^2}{\sum_i (\sigma_{hp_i} - \bar{\sigma}_{hp})^2}, \quad (2)$$

$$MSE = \frac{1}{n} \sum_{i=1}^n (\sigma_{ht_i} - \sigma_{hp_i})^2, \quad (3)$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (\sigma_{ht_i} - \sigma_{hp_i})^2}, \quad (4)$$

$$MAE = \frac{1}{n} \sum_{i=1}^n |\sigma_{ht_i} - \sigma_{hp_i}|, \quad (5)$$

$$MAPE = \left[\frac{1}{n} \sum_{i=1}^n \left| \frac{\sigma_{ht_i} - \sigma_{hp_i}}{\sigma_{ht_i}} \right| \right] \times 100, \quad (6)$$

em que σ_{ht} , σ_{hp} são, respectivamente os valores observados e estimados, n é o número de observações, $\bar{\sigma}_{ht}$, $\bar{\sigma}_{hp}$ valores médios previsto.

3.8 Análise de Explicabilidade dos Modelos (SHAP)

A explicabilidade dos modelos foi avaliada por meio do método *SHapley Additive exPlanations* (SHAP), fundamentado na teoria dos valores de Shapley, oriunda da teoria dos jogos cooperativos (Shapley, 1953). O método atribui a cada variável preditora uma contribuição marginal para a predição de cada observação, calculada com base em todas as possíveis combinações de subconjuntos de variáveis. Os valores SHAP satisfazem propriedades matemáticas como eficiência, simetria e aditividade, o que os torna uma abordagem consistente para a interpretação de modelos complexos (Lundberg; Lee, 2017). Diferente das métricas tradicionais de importância de variáveis, os valores SHAP permitem interpretar tanto local quanto global. Localmente, é possível identificar quais variáveis influenciaram a predição de cada observação e em que direção. Globalmente, a média dos valores absolutos de

SHAP fornece uma estimativa da importância média de cada variável ao longo do conjunto de dados. Essa característica torna o método adequado para análise de modelos em dados ambientais, onde as relações entre variáveis tendem a ser não lineares (Reichstein et al., 2019). Para os modelos com estrutura de árvore (*XGB*, *RF* e *DT*), os valores *SHAP* foram calculados por meio do pacote *treeshap* que implementa o algoritmo *TreeSHAP* proposto por Lundberg et al. (2020). Para os modelos *SVM* e *RNA*, que não possuem estrutura de árvore, os valores foram estimados por meio do pacote *fastshap* (Greenwell, 2023), que utiliza permutações aleatórias para aproximar os valores de *Shapley*. As visualizações foram geradas com o pacote *shapviz*. Nos gráficos do tipo *beeswarm*, cada ponto representa uma observação do conjunto de teste, o eixo horizontal indica a magnitude e direção do impacto da variável sobre a predição da FCO_2 , e a coloração representa o valor original da variável, do roxo (baixo) ao amarelo (alto).

4 RESULTADOS E DISCUSSÕES

4.1 Otimização de hiperparâmetros

Para garantir o melhor desempenho dos modelos, diferentes configurações de hiperparâmetros foram testadas e avaliadas com base na métrica de erro RMSE (Figura 1). No modelo de *RNA*, foram ajustados três principais hiperparâmetros: *hidden_units* (número de neurônios na camada oculta), *penalty* (penalidade de regularização) e *epochs* (número de épocas de treinamento). As combinações de hiperparâmetros testadas e seus respectivos valores de RMSE estão apresentados na Tabela 1. A regularização dos hiperparâmetros para o modelo de *RNA* indicou como melhor configuração 23 neurônios ocultos, uma penalidade de 100 e 77 épocas de treinamento (para o modelo de *RNA*), resultando em RMSE de $1,2881 \mu\text{mol m}^{-2} \text{s}^{-1}$, sendo esta configuração a que apresentou melhor desempenho preditivo, portanto, a menor margem de erro entre todas as combinações testadas (Figura 1 e Tabela 1). Em estudo de modelagem da emissão de CO_2 do solo em áreas de floresta tropical no Brasil, Canteral et al. (2023), observaram valores de RMSE associados ao processo de treinamento de RNAs (neuro-fuzzy, função e base radial e multilayer-perceptron) entre $1,04 \mu\text{mol m}^{-2} \text{s}^{-1}$ e $1,63 \mu\text{mol m}^{-2} \text{s}^{-1}$, valores estes próximos àqueles encontrados no presente estudo.

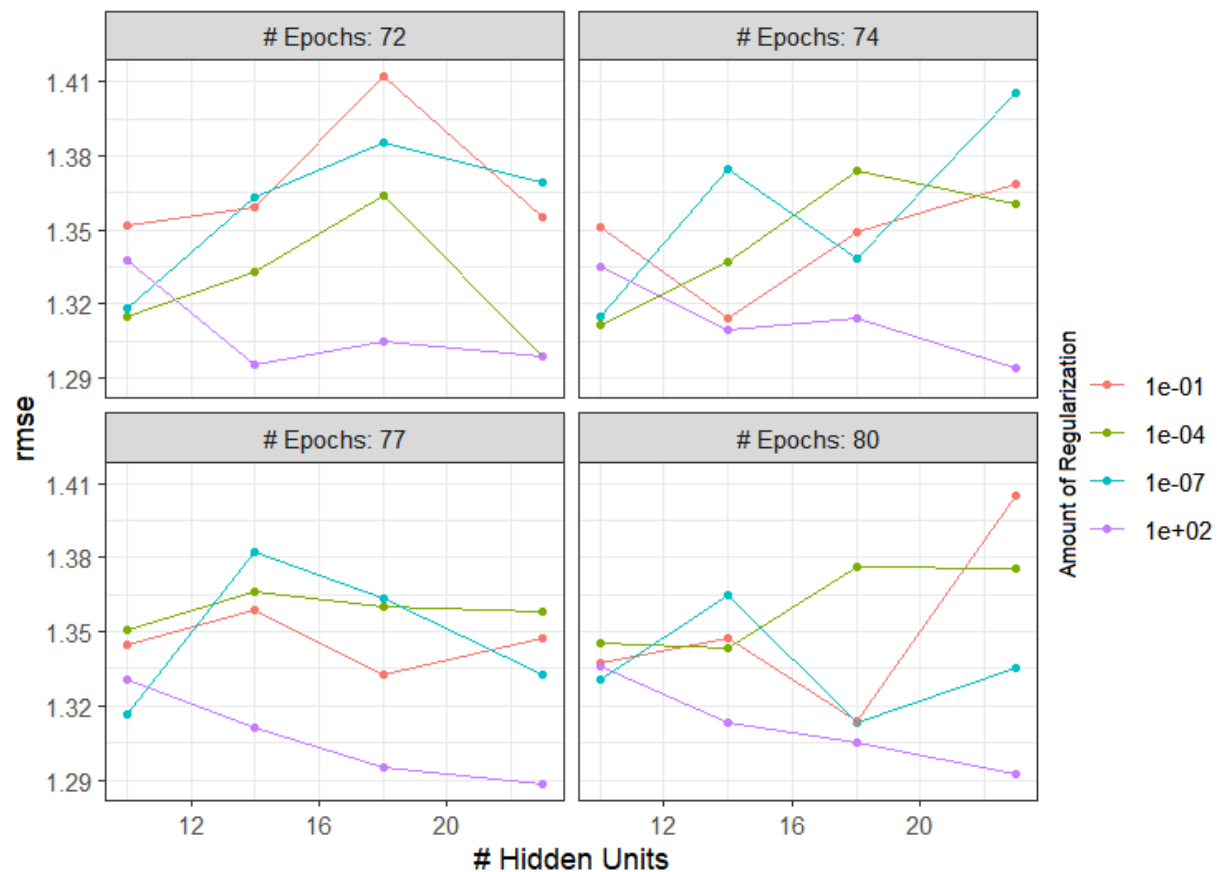


Figura 1 - Otimização de hiperparâmetros durante o processo de treinamento dos modelos para Redes Neurais Artificiais para a emissão de CO₂ do solo em áreas agrícolas (RMSE - $\mu\text{mol m}^{-2} \text{s}^{-1}$).

Tabela 1 - Rankeamento dos 10 melhores hiperparâmetros na etapa de treinamento dos modelos de aprendizado de máquina para as Redes Neurais Artificiais (RNA).

<i>hidden_units</i> *	<i>penalty</i> *	<i>epochs</i> *	RMSE
23	100	77	1,2881
23	100	80	1,2926
23	100	74	1,2942
18	100	77	1,2952
14	100	72	1,2955
23	100	72	1,2989
23	100	72	1,2990
18	100	72	1,3048
18	100	80	1,3054
14	100	74	1,3097

**hidden_units* - número de neurônios na camada oculta; *penalty*-penalidade de regularização; *epochs*-número de épocas de treinamento.

No modelo de *RF*, foram ajustados três principais hiperparâmetros: *mtry* (número de variáveis consideradas em cada divisão de nó), *trees* (número de árvores na floresta) e *min_n* (número mínimo de observações por nó terminal). As combinações de hiperparâmetros testadas e seus respectivos valores de RMSE estão apresentados na Tabela 2. A regularização dos hiperparâmetros para o modelo de *RF* indicou que a melhor configuração incluiu 341 árvores, *mtry* = 6 e *min_n* = 10, resultando em um RMSE de $1,1648 \mu\text{mol m}^{-2} \text{s}^{-1}$. Essa configuração foi a que apresentou o menor erro de predição, garantindo a maior precisão entre todas as combinações testadas (Figura 2 e Tabela 2). Estudos recentes demonstram que o modelo *RF* é uma abordagem confiável e amplamente aplicada para a previsão de emissões de CO_2 do solo. Por exemplo, Hamrani et al. (2020) compararam o desempenho de nove modelos de aprendizado de máquina e concluíram que o *RF* foi eficaz para capturar variações cíclicas e sazonais de emissões de CO_2 e N_2O em solos agrícolas. Neste estudo foram observados valores de RMSE baixos, em relação aos aqui apresentados e grande estabilidade nas previsões, se mostrando inferior apenas ao modelo de Memória de Longo e Curto Prazo (*Long short-term memory network-LSTM*). Em estudo recente, Lu et al. (2023) observaram que modelo *RF* para prever fluxos de respiração de ecossistema em plantações de trigo, onde o *RF* apresentou bom desempenho na integração de dados meteorológicos e características do solo para previsões em grande escala, com RMSE entre 1,12 e $1,13 \mu\text{mol m}^{-2} \text{s}^{-1}$.

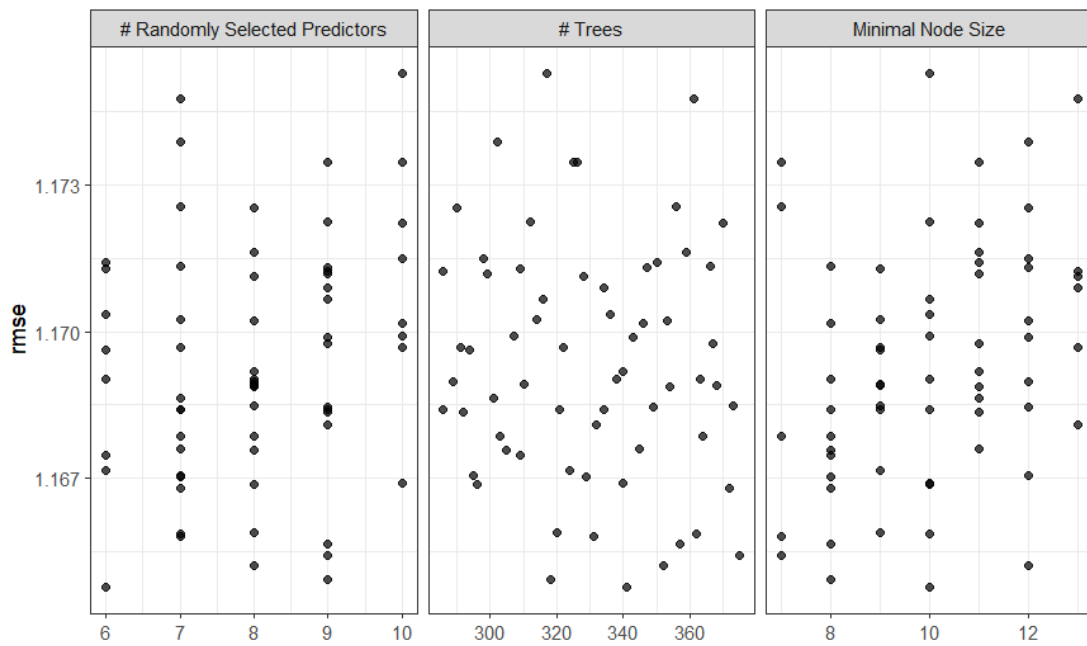


Figura 2 - Otimização de hiperparâmetros durante o processo de treinamento dos modelos para *Random Forest* para a emissão de CO₂ do solo em áreas agrícolas (RMSE - $\mu\text{mol m}^{-2} \text{s}^{-1}$).

Tabela 2 - Ranqueamento dos 10 melhores hiperparâmetros na etapa de treinamento dos modelos de aprendizado de máquina para as *Random Forest* (RF).

<i>mtry</i> *	<i>trees</i> *	<i>min_n</i> *	RMSE
6	341	10	1,1648
9	318	8	1,1649
8	352	12	1,1652
9	375	7	1,1654
9	357	8	1,1657
7	331	7	1,1658
7	362	10	1,1659
8	320	9	1,1659
7	372	8	1,1668
8	296	10	1,1669

* *mtry* - número de variáveis consideradas em cada divisão de nó; *trees* - número de árvores na floresta; *min_n* - número mínimo de observações por nó terminal.

No modelo de *SVM*, foram ajustados três principais hiperparâmetros: *cost* (parâmetro de penalização aplicado a erros de classificação), *rbf_sigma* (largura da função de base radial utilizada no kernel RBF) e *margin* (largura da margem de separação entre as observações). As combinações testadas e seus respectivos valores de RMSE encontram-se apresentados na Tabela 3. A regularização dos

hiperparâmetros indicou que a melhor configuração foi obtida com $cost = 1$, $rbf_sigma = 0,050$ e $margin = -0,01$, resultando em um RMSE de $1,2137 \mu\text{mol m}^{-2} \text{s}^{-1}$, representando o menor erro de predição entre todas as combinações avaliadas. Essa configuração demonstrou boa capacidade de generalização. De modo geral, a configuração ótima ($cost = 1$, $rbf_sigma = 0,050$, $margin = -0,01$) apresentou desempenho satisfatório, com erro moderado e boa estabilidade preditiva, destacando-se em relação às demais combinações. Embora o RMSE observado tenha sido superior ao obtido pelo modelo de *Random Forest* ($RF = 1,1648 \mu\text{mol m}^{-2} \text{s}^{-1}$), o SVM mostrou-se eficiente na predição dos emissão de CO_2 do solo, especialmente na modelagem de padrões não lineares complexos.

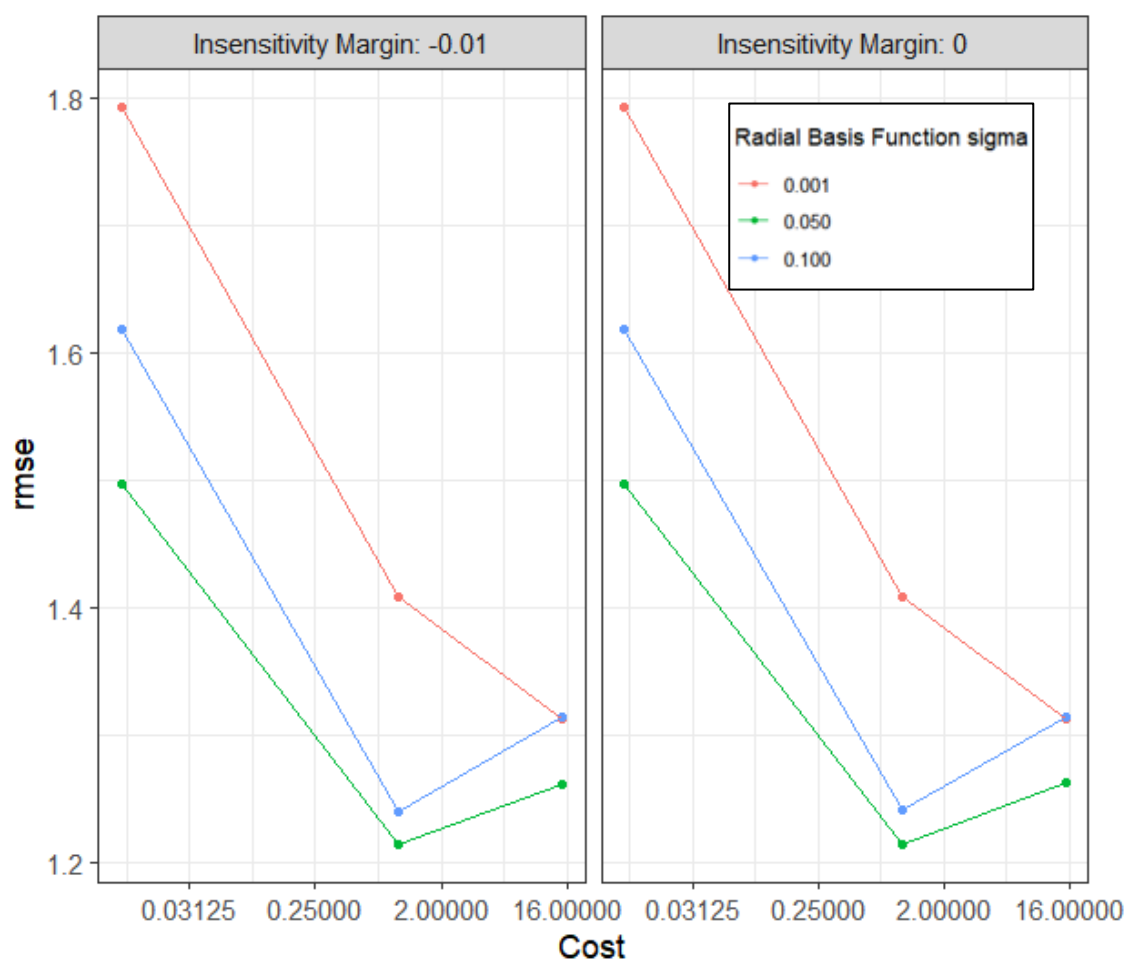


Figura 3 - Otimização de hiperparâmetros durante o processo de treinamento dos modelos para *Support Vector Machine* para a emissão de CO_2 do solo em áreas agrícolas (RMSE - $\mu\text{mol m}^{-2} \text{s}^{-1}$).

Tabela 3 - Rankeamento dos 10 melhores hiperparâmetros na etapa de treinamento dos modelos de aprendizado de máquina para *Support Vector Machine* (SVM).

<i>cost</i> *	<i>rbf_sigma</i> *	<i>margin</i> *	RMSE
1	0,050	-0,01	1,2137
1	0,050	0,00	1,2141
1	0,100	-0,01	1,2403
1	0,100	0,00	1,2408
15	0,050	-0,01	1,2613
15	0,050	0,00	1,2628
15	0,001	-0,01	1,3128
15	0,001	0,00	1,3133
15	0,100	0,00	1,3146
15	0,100	-0,01	1,3147

* *cost* - parâmetro de penalização aplicado a erros de classificação; *rbf_sigma* - largura da função de base radial; *margin* - largura da margem de separação entre as observações.

Para o ajuste do modelo de *Decision Tree* (DT), diferentes combinações de hiperparâmetros foram avaliadas com base na métrica de erro RMSE, conforme apresentado na Tabela 4. Os principais hiperparâmetros ajustados foram: *cost_complexity*, que controla o grau de poda da árvore e evita que o modelo se torne excessivamente complexo; *tree_depth*, que determina a profundidade máxima da árvore e, portanto, sua capacidade de capturar relações complexas; e *min_n*, que define o número mínimo de observações necessárias em cada nó terminal. A regularização dos hiperparâmetros indicou que a melhor configuração obtida foi com *cost_complexity* = 1,00e-08, *tree_depth* = 26 e *min_n* = 50, resultando em um RMSE de 1,3197 $\mu\text{mol m}^{-2} \text{s}^{-1}$, sendo este o menor erro entre todas as combinações testadas. Configurações com valores próximos para *cost_complexity* (2,37e-07 e 5,62e-06) apresentaram desempenho similar, indicando baixa sensibilidade do modelo a variações nesse parâmetro dentro da faixa testada. A persistência da profundidade da árvore para *tree_depth* = 26 sugere que maiores profundidades não contribuem para ganhos expressivos de acurácia e podem elevar o risco de sobreajuste. Além disso, a manutenção de *min_n* = 50 em todas as melhores configurações indica que um número mínimo mais elevado de observações nos nós terminais favoreceu a generalização do modelo, reduzindo oscilações e instabilidade preditiva. Ainda que o modelo de *Decision Tree* tenha apresentado desempenho consistente, o RMSE observado foi superior ao obtido pelos modelos *Random Forest* (RF = 1,1648 $\mu\text{mol m}^{-2} \text{s}^{-1}$) e SVM (1,2137 $\mu\text{mol m}^{-2} \text{s}^{-1}$), indicando menor capacidade de capturar padrões complexos nos dados. Essa diferença é esperada, visto que o DT representa uma

estrutura mais simples e sensível a variações nos dados, enquanto modelos *ensemble* e métodos baseados em kernel tendem a apresentar maior robustez e desempenho preditivo superior.

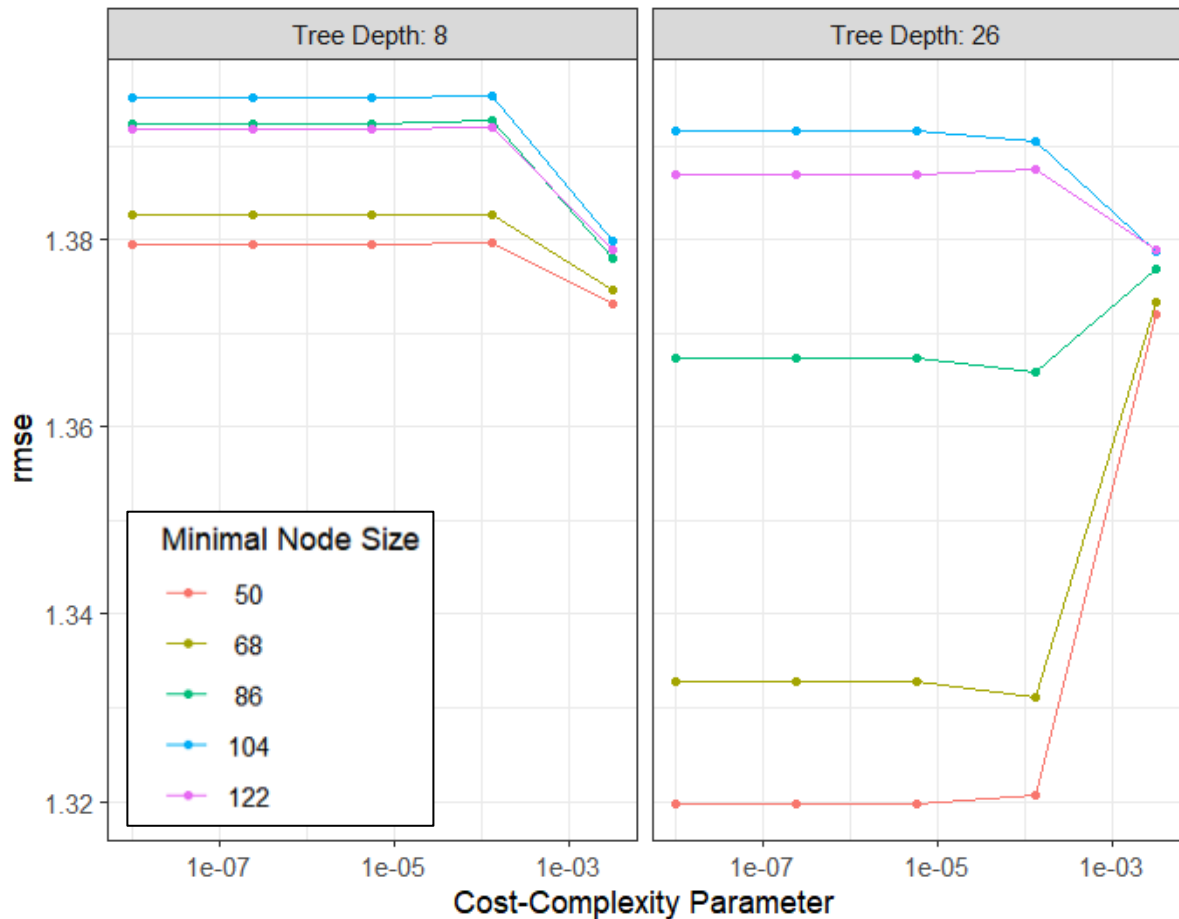


Figura 4 - Otimização de hiperparâmetros durante o processo de treinamento dos modelos para Decision Tree para a emissão de CO₂ do solo em áreas agrícolas (RMSE - $\mu\text{mol m}^{-2} \text{s}^{-1}$).

Tabela 4 - Ranqueamento dos 10 melhores hiperparâmetros na etapa de treinamento dos modelos de aprendizado de máquina para *Decision Tree* (DT).

<i>cost_complexity</i> *	<i>tree_depth</i> *	<i>min_n</i> *	RMSE
1,00 e-08	26	50	1,3197
2,37 e-07	26	50	1,3197
5,62 e-06	26	50	1,3197
1,33 e-04	26	50	1,3206
1,33 e-04	26	68	1,3311
1,00 e-08	26	68	1,3327
2,37 e-07	26	68	1,3327
5,62 e-06	26	68	1,3327

1,33 e-04	26	86	1,3658
1,00 e-08	26	86	1,3674

* *cost_complexity* – grau de poda da árvore; *tree_depth* – profundidade máxima da árvore; *min_n* – número mínimo de observações.

Para garantir o melhor desempenho do modelo *XGB*, diferentes combinações de hiperparâmetros foram testadas e avaliadas com base na métrica de erro RMSE, de forma semelhante ao procedimento adotado para os demais modelos. O processo de otimização foi conduzido em etapas sequenciais, visando refinar progressivamente os principais parâmetros do modelo. Na primeira etapa, foram ajustados os hiperparâmetros taxa de aprendizado (*learn_rate*) e número de árvores (*trees*) (Figura 5 e Tabela 5). As melhores combinações iniciais apresentaram valores de RMSE em torno de $1,28 \mu\text{mol m}^{-2} \text{s}^{-1}$, indicando que o aumento gradual do número de árvores associado a uma taxa de aprendizado baixa contribuiu para maior estabilidade do modelo.

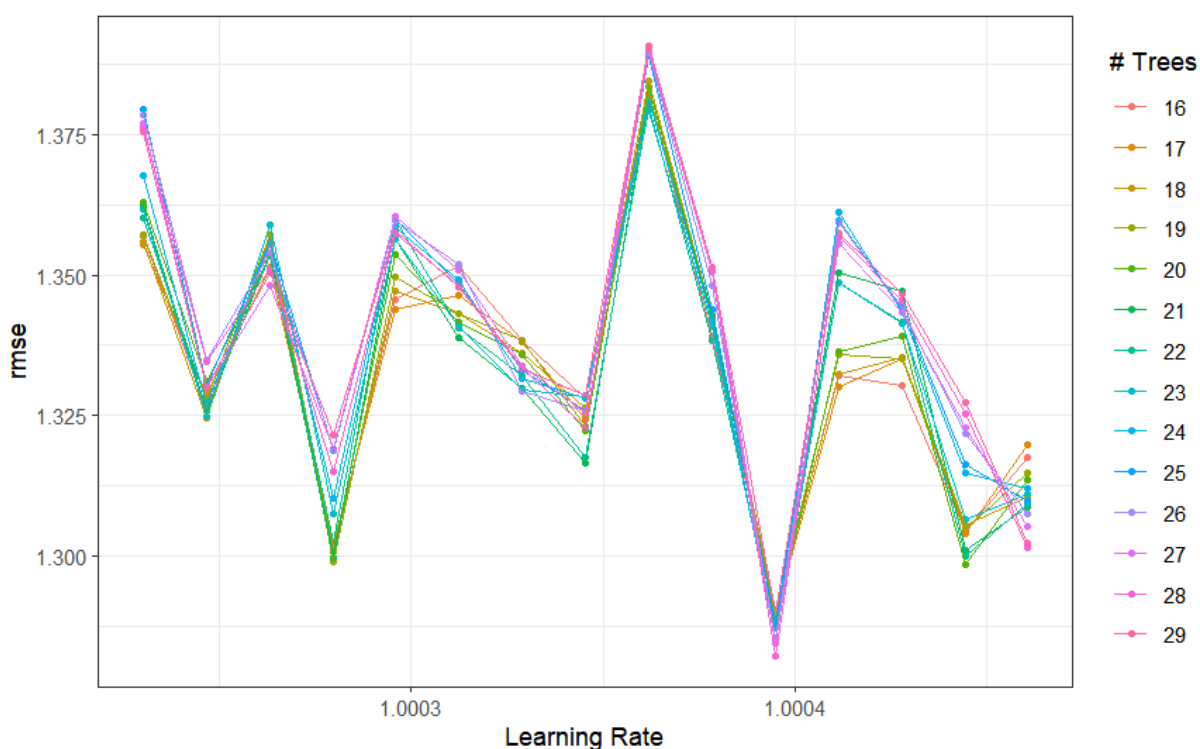


Figura 5 – Primeira otimização de hiperparâmetros durante o processo de treinamento dos modelos para *XGB* para a emissão de CO_2 do solo em áreas agrícolas (RMSE - $\mu\text{mol m}^{-2} \text{s}^{-1}$).

Tabela 5 - Rankeamento dos 10 melhores hiperparâmetros na etapa 1 de treinamento dos modelos de aprendizado de máquina para *Extreme Gradient Boosting* (XGB).

<i>learn_rate</i> *	<i>trees</i> *	RMSE
1,000395	28	1,2822
1,000395	27	1,2843
1,000395	24	1,2851
1,000395	26	1,2852
1,000395	23	1,2855
1,000395	22	1,2871
1,000395	25	1,2876
1,000395	21	1,2881
1,000395	20	1,2887
1,000395	29	1,2899

* *learn_rate* – taxa de aprendizado; *trees* - número de árvores na floresta.

Na segunda etapa, foram ajustados os parâmetros estruturais das árvores: profundidade máxima (*tree_depth*) e número mínimo de observações por nó (*min_n*) (Figura 6 e Tabela 6). A melhor combinação encontrada foi *tree_depth* = 10 e *min_n* = 222, resultando em $RMSE = 1,2333 \mu\text{mol m}^{-2} \text{s}^{-1}$, representando melhora significativa no desempenho em relação à etapa anterior. Esse resultado indica que árvores mais profundas, combinadas a um controle do tamanho mínimo dos nós, permitiram capturar relações não lineares mais complexas entre as variáveis ambientais e as emissões de CO₂ do solo.

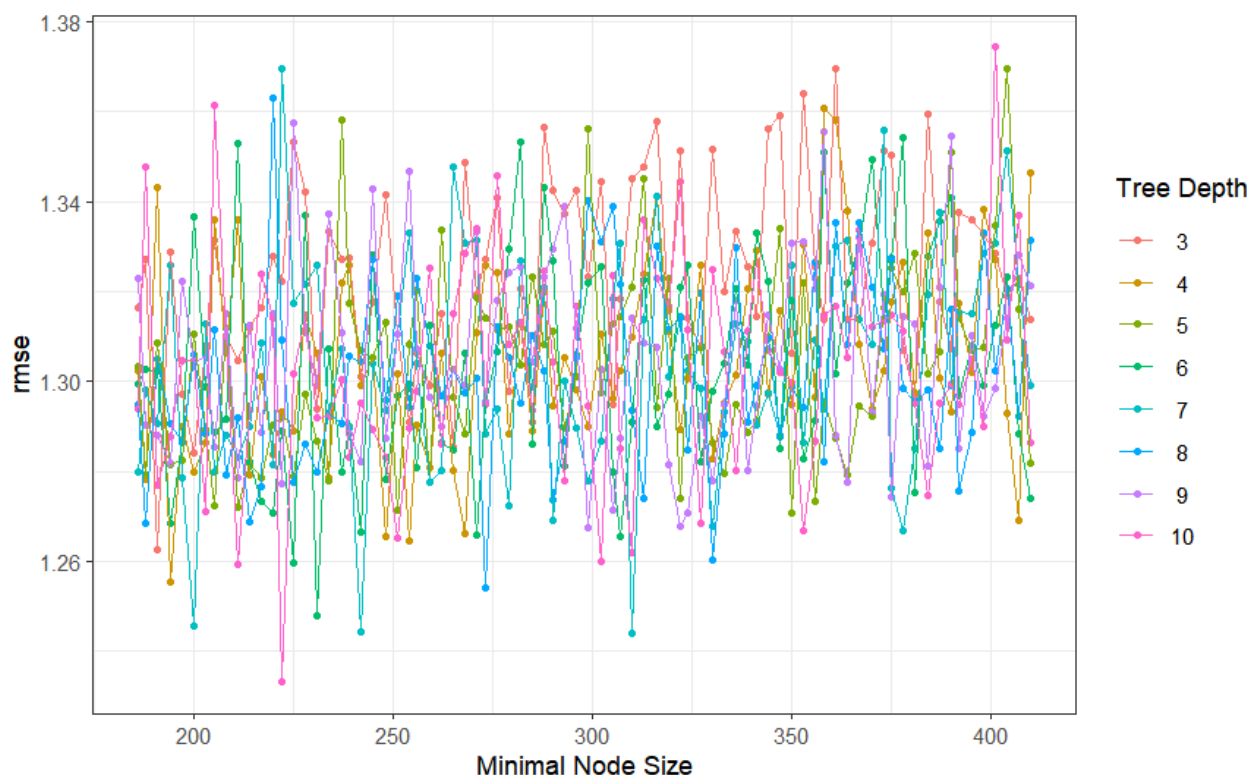


Figura 6 – Segunda otimização de hiperparâmetros durante o processo de treinamento dos modelos para *XGB* para a emissão de CO_2 do solo em áreas agrícolas (RMSE - $\mu\text{mol m}^{-2} \text{s}^{-1}$).

Tabela 6 - Rankeamento dos 10 melhores hiperparâmetros na etapa 2 de treinamento dos modelos de aprendizado de máquina para *Extreme Gradient Boosting (XGB)*.

<i>min_n</i> *	<i>tree_depth</i> *	RMSE
222	10	1,2333
310	7	1,2440
242	7	1,2443
200	7	1,2456
231	6	1,2480
273	8	1,2542
194	4	1,2556
211	10	1,2595
225	6	1,2596
302	10	1,2600

* *min_n* – número mínimo de observações; *tree_depth* – profundidade máxima da árvore.

Na terceira etapa, foi avaliado o hiperparâmetro *loss_reduction (gamma)*, responsável por controlar a redução mínima da função de perda necessária para realizar novas divisões nos nós (Figura 7 e Tabela 7). Os valores testados resultaram

em RMSE próximo de $1,2532 \mu\text{mol m}^{-2} \text{s}^{-1}$, indicando que esse parâmetro teve influência moderada no desempenho do modelo dentro da faixa analisada.

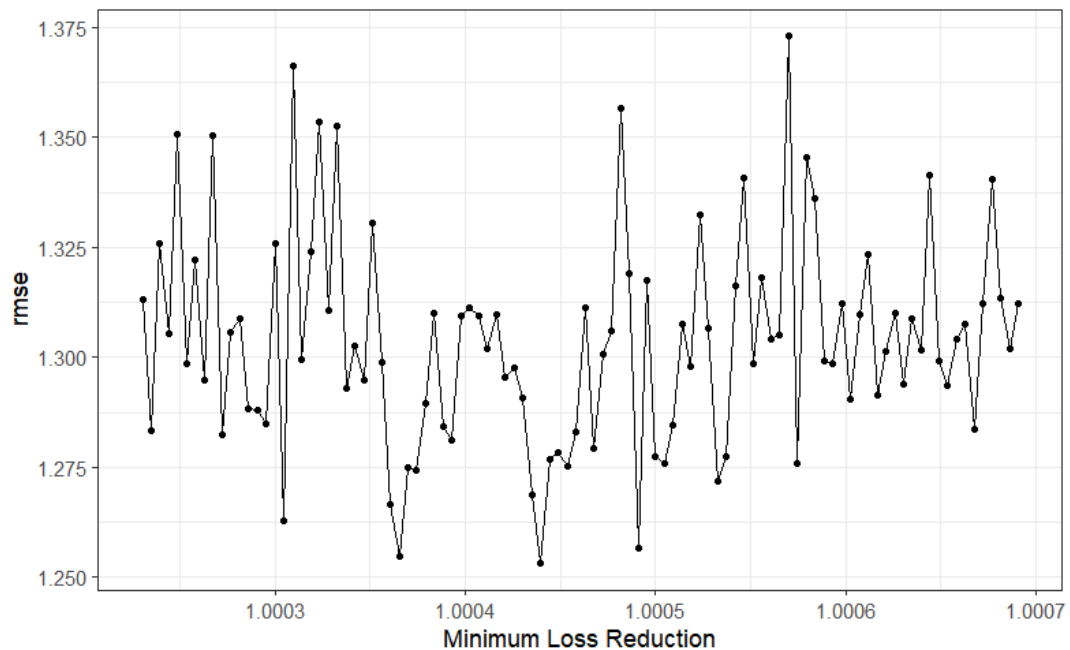


Figura 7 – Terceira otimização de hiperparâmetros durante o processo de treinamento dos modelos para XGB para a emissão de CO_2 do solo em áreas agrícolas (RMSE - $\mu\text{mol m}^{-2} \text{s}^{-1}$).

Tabela 7 - Ranqueamento dos 10 melhores hiperparâmetros na etapa 3 de treinamento dos modelos de aprendizado de máquina para *Extreme Gradient Boosting* (XGB).

<i>loss_reduction</i> *	RMSE
1,000440	1,2532
1,000365	1,2547
1,000491	1,2565
1,000305	1,2630
1,000361	1,2665
1,000435	1,2689
1,000533	1,2717
1,000375	1,2743
1,000370	1,2749
1,000454	1,2753

* *loss_reduction* – redução de perda mínima para divisão.

Na quarta etapa, foram ajustados os parâmetros de amostragem: *sample_size* (fração das observações usadas em cada árvore) e *mtry* (fração de variáveis candidatas por divisão) (Figura 8 e Tabela 8). A melhor configuração foi *sample_size* = 0,98 e *mtry* = 0,70, com RMSE = $1,2309 \mu\text{mol m}^{-2} \text{s}^{-1}$, sugerindo que a introdução

de aleatoriedade controlada contribuiu para reduzir o sobreajuste e aumentar a capacidade de generalização do modelo.

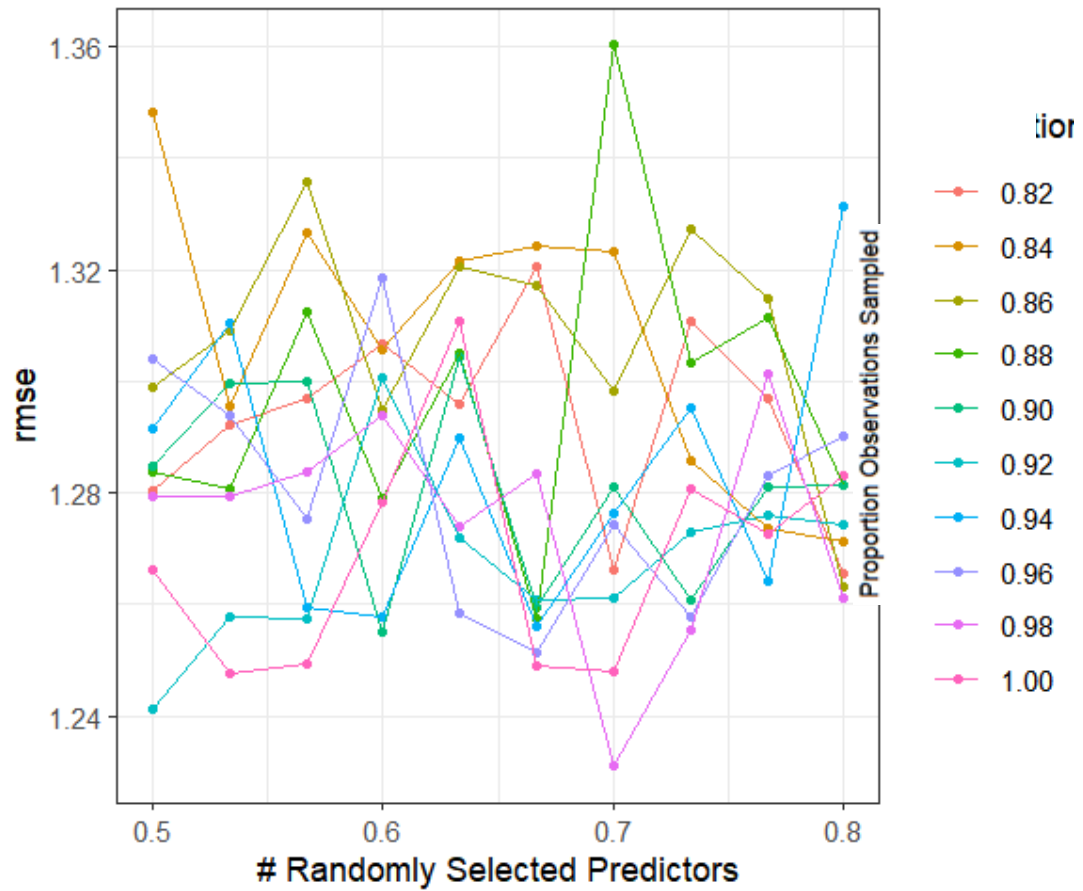


Figura 8 – Quarta otimização de hiperparâmetros durante o processo de treinamento dos modelos para *XGB* para a emissão de CO_2 do solo em áreas agrícolas (RMSE - $\mu\text{mol m}^{-2} \text{s}^{-1}$).

Tabela 8 - Rankeamento dos 10 melhores hiperparâmetros na etapa 4 de treinamento dos modelos de aprendizado de máquina para *Extreme Gradient Boosting (XGB)*.

<i>mtry</i> *	<i>sample_size</i> *	RMSE
0,70	0,98	1,2309
0,50	0,92	1,2411
0,53	1,00	1,2475
0,70	1,00	1,2478
0,67	1,00	1,2489
0,57	1,00	1,2492
0,67	0,96	1,2512
0,60	0,90	1,2552
0,73	0,98	1,2555
0,67	0,94	1,2561

* *mtry* - número de variáveis consideradas em cada divisão de nó; *sample_size* - fração das observações usadas em cada árvore.

Por fim, na etapa de ajuste fino, foram novamente testadas combinações de *learn_rate* e *trees* em faixas mais amplas (Figura 9 e Tabela 9). A melhor configuração geral do modelo foi obtida com *trees* = 700 e *learn_rate* = 0,24, resultando em RMSE = 1,1638 $\mu\text{mol m}^{-2} \text{s}^{-1}$, o menor erro observado durante todo o processo de calibração. De modo geral, a calibração mostrou que o desempenho do *XGB* é fortemente dependente do equilíbrio entre número de árvores e taxa de aprendizado, além do controle da complexidade estrutural das árvores.

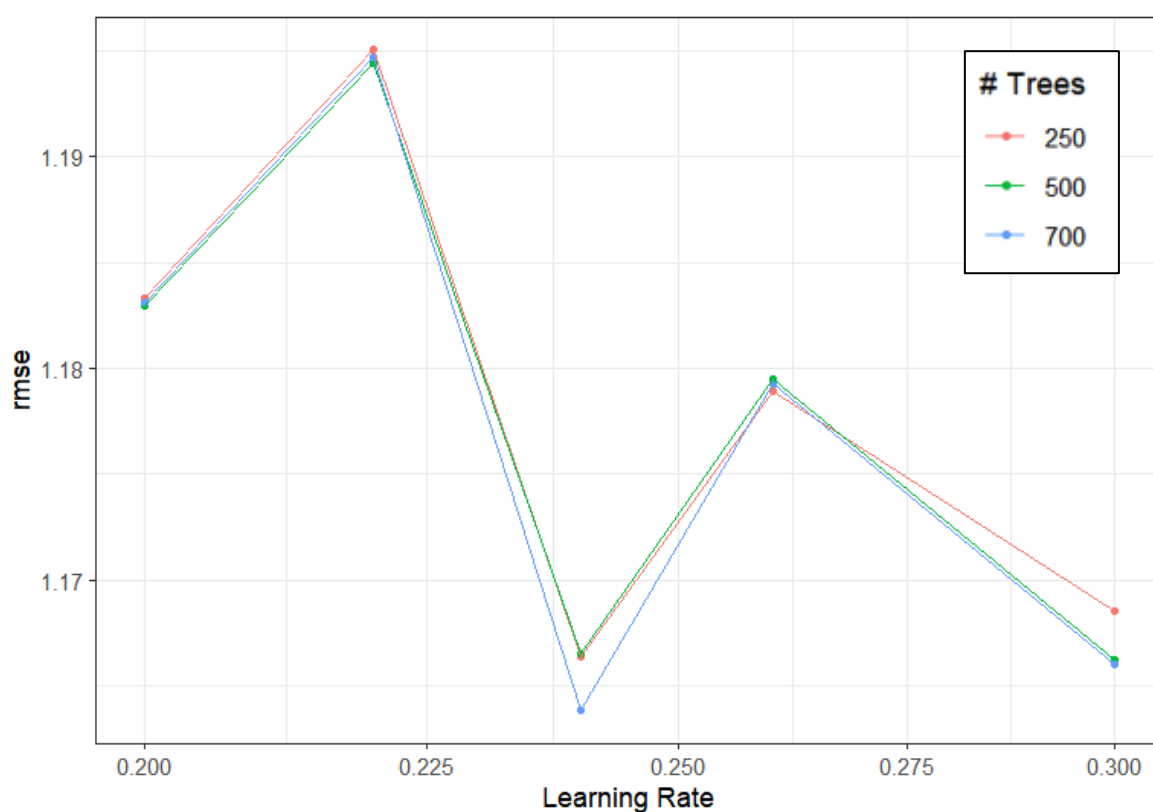


Figura 9 – Quinta otimização de hiperparâmetros durante o processo de treinamento dos modelos para *XGB* para a emissão de CO_2 do solo em áreas agrícolas (RMSE - $\mu\text{mol m}^{-2} \text{s}^{-1}$).

Tabela 9 - Ranqueamento dos 10 melhores hiperparâmetros na etapa 5 de treinamento dos modelos de aprendizado de máquina para *Extreme Gradient Boosting (XGB)*.

<i>learn_rate</i> *	<i>trees</i> *	RMSE
0,24	700	1,1638

0,30	700	1,1660
0,30	500	1,1662
0,24	250	1,1663
0,24	500	1,1665
0,30	250	1,1685
0,26	250	1,1789
0,26	700	1,1793
0,26	500	1,1795
500	0,20	1.1830

* *learn_rate* – taxa de aprendizado; *trees* - número de árvores na floresta.

4.2 Desempenho do Modelo

O desempenho preditivo dos modelos de aprendizado de máquina foi avaliado com base no conjunto de teste, utilizando diferentes métricas estatísticas (Tabela 10 e Figura 10). De forma geral, o modelo *Random Forest (RF)* apresentou o melhor desempenho entre os algoritmos avaliados. O *RF* obteve o maior coeficiente de correlação ($r = 0,83$) e o maior coeficiente de determinação ($R^2 = 0,69$), indicando maior capacidade de explicar a variabilidade das emissões de CO_2 do solo. Além disso, apresentou os menores valores de erro, com $\text{RMSE} = 1,18 \mu\text{mol m}^{-2} \text{s}^{-1}$ e $\text{MAE} = 0,76$, evidenciando maior precisão e menor dispersão dos resíduos. O modelo *XGB* apresentou desempenho muito próximo ao *RF*, destacando-se como o segundo melhor modelo. O *XGB* obteve $r = 0,82$ e $R^2 = 0,66$, com $\text{RMSE} = 1,21 \mu\text{mol m}^{-2} \text{s}^{-1}$. Esses resultados demonstram elevada capacidade de generalização e reforçam a eficiência de métodos baseados em *boosting* na modelagem de processos ambientais complexos. O desempenho semelhante ao *RF* indica que ambos os modelos *ensemble* foram mais eficazes na captura de interações não lineares entre variáveis climáticas e atributos do solo. O modelo *Support Vector Machine (SVM)* apresentou desempenho intermediário, com $R^2 = 0,64$ e $\text{RMSE} = 1,25 \mu\text{mol m}^{-2} \text{s}^{-1}$. Embora tenha apresentado erro ligeiramente superior aos modelos baseados em árvores, o *SVM* mostrou boa capacidade de modelar relações não lineares, mantendo desempenho estável. As Redes Neurais Artificiais (*RNA*) apresentaram desempenho semelhante ao da *Decision Tree*, com $R^2 = 0,61$ e $\text{RMSE} = 1,32 \mu\text{mol m}^{-2} \text{s}^{-1}$. Apesar de conseguirem capturar padrões complexos, o desempenho inferior aos modelos *ensemble* pode estar associado à maior sensibilidade a ruídos e à necessidade de maior volume de dados para calibração mais robusta. O modelo de *Decision Tree (DT)* apresentou o menor desempenho entre os avaliados ($R^2 = 0,61$; $\text{RMSE} = 1,30 \mu\text{mol m}^{-2} \text{s}^{-1}$), resultado esperado devido à sua estrutura mais simples e maior

suscetibilidade a variações nos dados. Modelos baseados em uma única árvore tendem a apresentar menor capacidade de generalização quando comparados a métodos *ensemble*, como *RF* e *XGB*. De modo geral, os resultados indicam que modelos *ensemble* baseados em árvores (*RF* e *XGB*) foram os mais eficientes para estimar a emissão de CO₂ do solo, apresentando melhor equilíbrio e robustez frente à complexidade dos dados ambientais.

Tabela 10 – Comparação dos indicadores de desempenho dos modelos.

Vetores de métrica	<i>RNA</i>	<i>RF</i>	<i>SVM</i>	<i>DT</i>	<i>XGB</i>
r	0,7787	0,8277	0,8027	0,7832	0,8155
R ²	0,6064	0,6851	0,6443	0,6134	0,6645
MSE	1,7332	1,3957	1,5711	1,6975	1,4729
RMSE	1,3165	1,1814	1,2534	1,3023	1,2136
MAE	0,8902	0,7628	0,7850	0,8736	0,7875
MAPE	47,2021	44,7157	39,3288	49,0000	42,2147

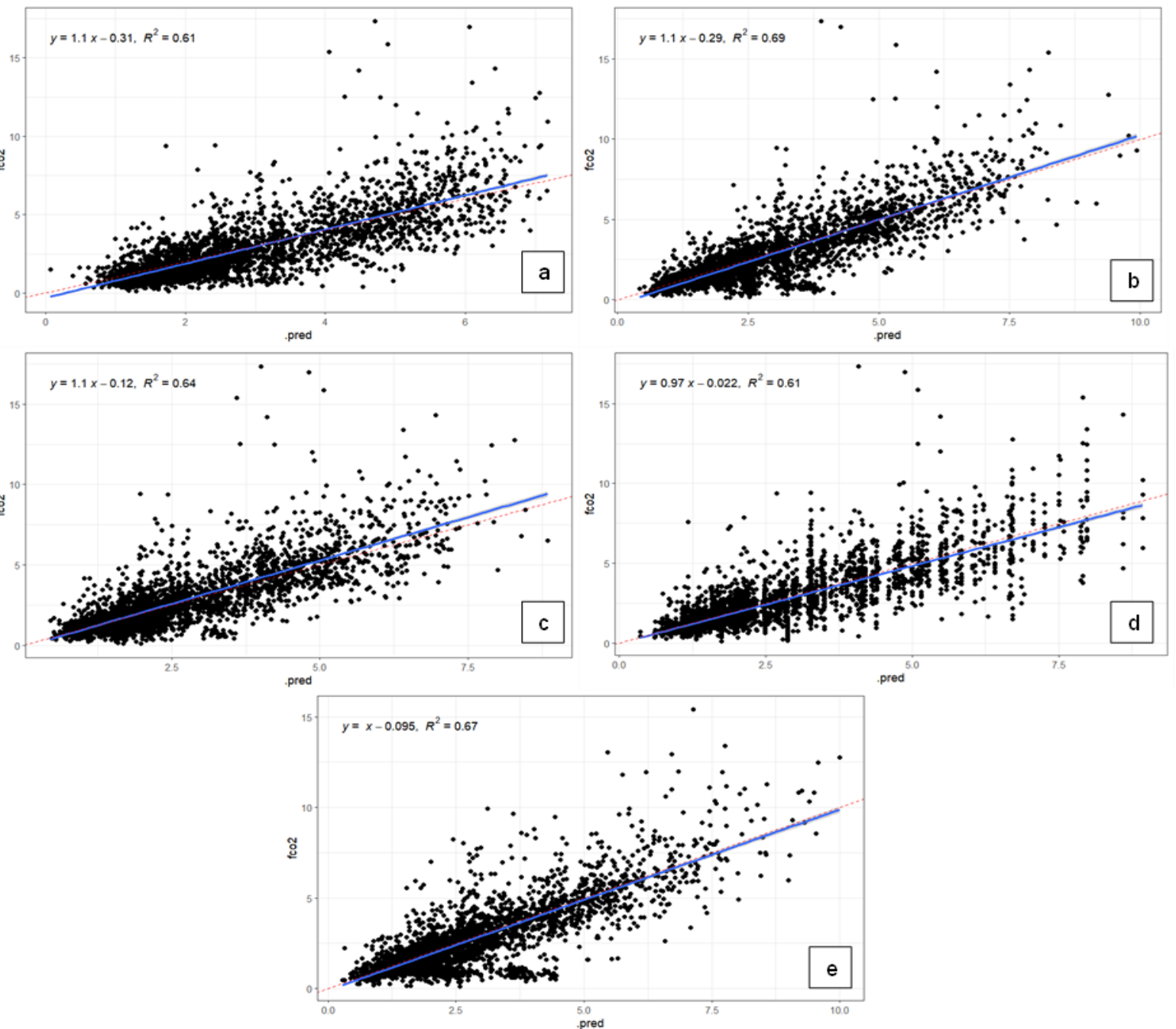


Figura 10 - Comparação do desempenho dos modelos na base de teste para: a) RNA, com um coeficiente de determinação de $R^2 = 0,61$; b) RF com coeficiente de determinação de $R^2 = 0,69$; c) SVM com coeficiente de determinação de $R^2 = 0,64$; d) DT com coeficiente de determinação de $R^2 = 0,61$; e) XGB com coeficiente de determinação de $R^2 = 0,67$.

4.3 Importância das variáveis para os modelos

Todos os modelos foram analisados quanto à importância das variáveis utilizadas para prever as emissões de CO₂ do solo (Figura 11). De forma geral, observou-se a predominância de variáveis climáticas, dinâmicas atmosféricas e indicadores da atividade da vegetação, além de alguns atributos químicos e físicos do solo, evidenciando o caráter multifatorial que controla a respiração do solo. No modelo de Redes Neurais Artificiais (RNA) (Figura 11a), as variáveis mais relevantes foram umidade do ar/solo (u_{med}), fator de impacto solar (SIF), magnésio (Mg), temperatura máxima (T_{max}), matéria orgânica (MO) e pressão de vapor do solo (Pk_{pa}). A forte influência de variáveis climáticas indica que o modelo foi sensível às condições térmicas e hídricas do ambiente. A umidade regula a disponibilidade de água e a difusão de oxigênio no solo, afetando diretamente a atividade microbiana. A temperatura máxima está associada ao aumento da atividade enzimática e da decomposição da matéria orgânica. Já a SIF reflete a atividade fotossintética da vegetação, que influencia indiretamente a respiração radicular e o aporte de carbono ao solo. No modelo de *Random Forest* (RF) (Figura 11b), destacaram-se SIF, concentração média de CO₂ atmosférico (xCO_2), temperatura mínima (T_{min}), umidade (u_{med}) e temperatura do solo (T_s). A importância da SIF sugere forte conexão entre a atividade fotossintética e as emissões de CO₂ do solo, uma vez que maior atividade vegetal implica maior aporte de exsudatos radiculares e maior respiração autotrófica. A temperatura mínima influencia a respiração microbiana ao controlar o metabolismo em períodos mais frios, enquanto a umidade regula as condições aeróbicas do solo. A presença de xCO_2 indica que o modelo também capturou relações entre a concentração atmosférica de carbono e os gradientes de difusão do gás entre solo e atmosfera. O modelo de *Support Vector Machine* (SVM) (Figura 11c) apresentou maior importância para variáveis químicas do solo, como pH e magnésio (Mg), além de SIF, umidade do solo (U_s) e umidade do ar (u_{med}). Isso sugere que o SVM foi particularmente sensível às condições químicas e de fertilidade do solo, que influenciam o crescimento radicular, a atividade microbiana e a decomposição da matéria orgânica. O pH, por exemplo, regula a disponibilidade de nutrientes e a atividade enzimática microbiana, enquanto nutrientes como Mg participam de processos metabólicos vegetais e microbianos que impactam a respiração do solo. Na *Decision Tree* (DT) (Figura 11d), observou-se maior destaque para SIF,

temperatura mínima ($T_{\min}$), $x\text{CO}_2$, temperatura média (T_{med}) e radiação fotossinteticamente ativa (PAR). Esse conjunto reforça o papel das variáveis climáticas e da atividade da vegetação na regulação das emissões de CO_2 . A radiação solar (PAR) controla a fotossíntese e, conseqüentemente, o fornecimento de carbono para o sistema solo-planta. Já as variáveis térmicas modulam a atividade biológica no solo, influenciando tanto a respiração microbiana quanto radicular. Por fim, no modelo *XGB* (Figura 11e), a variável de maior importância foi SIF, seguida por umidade (U_{med}), temperatura mínima ($T_{\min}$) e $x\text{CO}_2$, além de variáveis químicas como Mg, CTC e SB. Esse padrão mostra que o *XGB* integrou de forma eficiente informações sobre atividade da vegetação, condições climáticas e fertilidade do solo. A capacidade de troca catiônica (CTC) e a saturação por bases (SB) estão relacionadas à retenção e disponibilidade de nutrientes, que afetam o crescimento radicular e a atividade microbiana, refletindo nas emissões de CO_2 . De maneira geral, a recorrência de variáveis como SIF, umidade e temperatura em praticamente todos os modelos confirma o papel central das condições climáticas e da atividade da vegetação na regulação da respiração do solo. Esses resultados são coerentes, pois a emissão de CO_2 do solo é resultado da interação entre respiração microbiana, respiração radicular e decomposição da matéria orgânica, processos dependentes de temperatura, disponibilidade hídrica e aporte de carbono proveniente da fotossíntese.

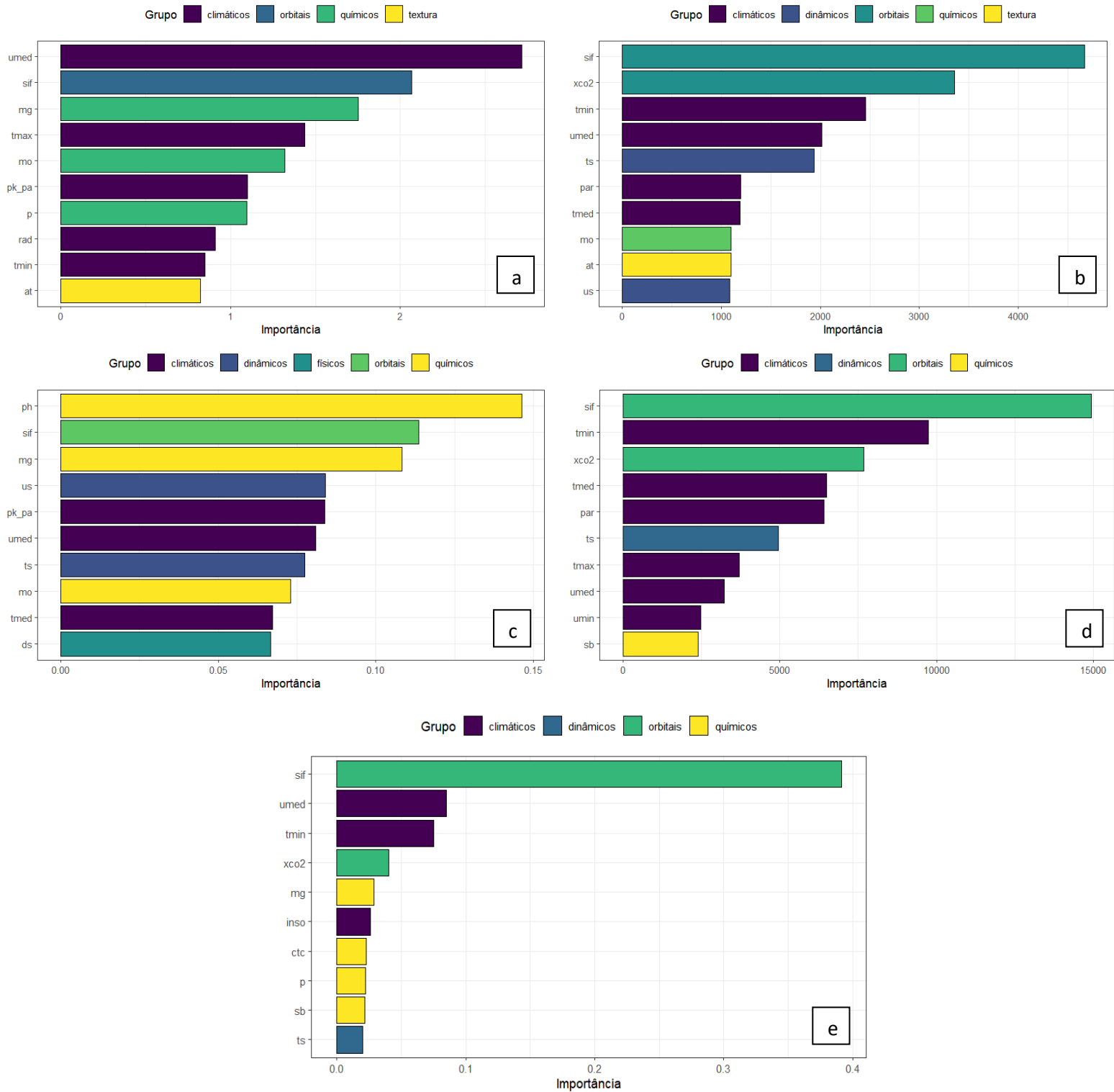


Figura 11 – a) importância relativa das variáveis para o modelo RNA; b) importância relativa das variáveis para o modelo RF; c) importância relativa das variáveis para o modelo SVM; d) importância relativa das variáveis para o modelo DT; e) importância relativa das variáveis para o modelo XGB.

4.4 Análise de Explicabilidade dos Modelos (SHAP)

A análise de explicabilidade dos modelos foi conduzida por meio do método *SHapley Additive exPlanations* (SHAP), uma abordagem fundamentada na teoria dos jogos cooperativos que permite quantificar a contribuição individual de cada variável preditora sobre as previsões geradas pelo modelo. Diferentemente das métricas tradicionais de importância de variáveis, que fornecem apenas uma medida global de relevância, os valores SHAP apresentam uma perspectiva local e global simultaneamente, possibilitando identificar não apenas quais variáveis são mais importantes, mas também a direção e a magnitude de seus efeitos sobre cada observação individual. Para a visualização dos resultados, foram utilizados os gráficos do tipo *beeswarm*, nos quais cada ponto representa uma observação, o eixo horizontal indica o valor SHAP (impacto positivo ou negativo sobre a previsão da FCO₂) e a coloração do ponto reflete o valor original da variável. Os resultados são apresentados para os cinco modelos avaliados: *Extreme Gradient Boosting* (XGB), *Random Forest* (RF), *Support Vector Machine* (SVM), *Decision Tree* (DT) e Redes Neurais Artificiais (RNA).

No modelo XGB (Figura 12), a fluorescência solar induzida (SIF) destacou-se como a variável de maior impacto, seguida pela concentração média de CO₂ atmosférico (xCO₂) e pela umidade relativa do ar (umed). Observa-se que valores elevados dessas três variáveis (representados pela coloração amarela e laranja) estão associados a valores SHAP positivos, indicando que condições de maior atividade fotossintética, maior concentração de CO₂ atmosférico e maior umidade tendem a elevar as previsões de FCO₂. O magnésio (mg) apresentou comportamento distinto, com valores baixos associados a fortes contribuições negativas, sugerindo que sua deficiência pode limitar a atividade biológica do solo. A amplitude dos valores SHAP para SIF e xCO₂ foi superior às demais variáveis, reforçando o papel dessas variáveis na modelagem da respiração do solo pelo XGB.

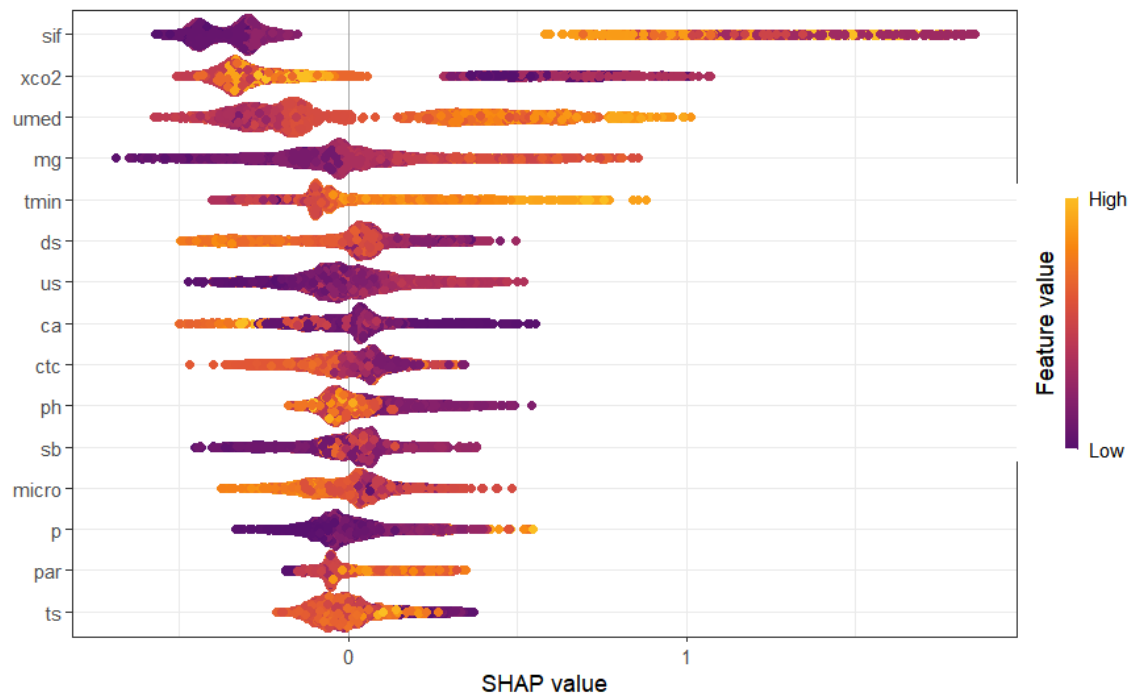


Figura 12 – Gráfico *beeswarm* dos valores SHAP para o modelo *XGB*.

No modelo *RF* (Figura 13), a distribuição dos valores SHAP apresentou padrão semelhante ao *XGB*, com predominância de *SIF*, *xCO₂* e *umed* como variáveis de maior influência. No entanto, o *RF* atribuiu maior relevância a variáveis térmicas, especialmente a temperatura mínima do ar (*tmin*) e a temperatura do solo (*ts*), além da temperatura média (*tmed*). Esse comportamento indica que o *RF* captou com maior sensibilidade a influência das condições térmicas sobre a respiração do solo, processo conhecido por apresentar relação exponencial com a temperatura. A distribuição mais homogênea dos valores SHAP entre as variáveis sugere que o *RF* distribui a importância preditiva de forma mais equilibrada, característica esperada em modelos ensemble que combinam múltiplas árvores de decisão independentes.

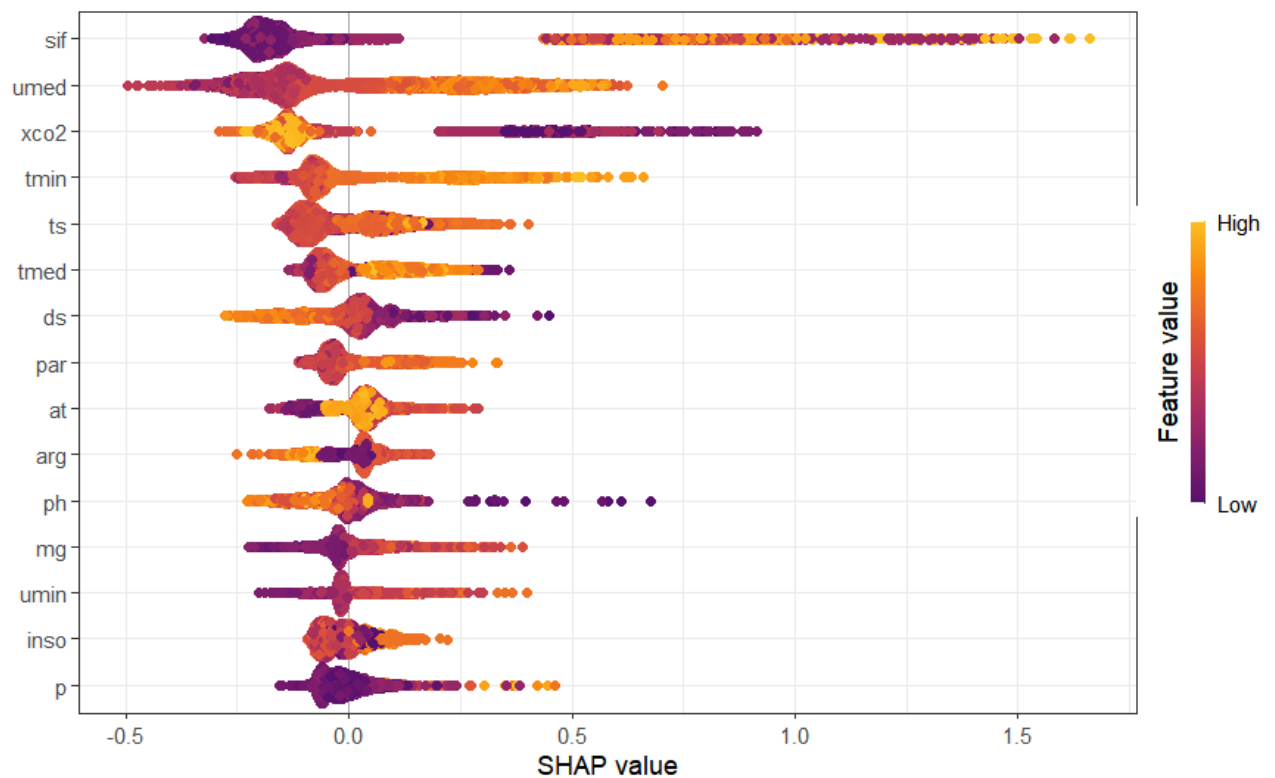


Figura 13 – Gráfico *beeswarm* dos valores SHAP para o modelo *RF*.

O modelo *SVM* (Figura 14) apresentou um perfil de explicabilidade distinto dos modelos baseados em árvores. A *SIF* manteve-se como variável de maior destaque, porém com menor amplitude de valores SHAP em comparação ao *XGB* e ao *RF*. O pH do solo emergiu como segundo preditor mais relevante, com valores elevados de pH associados a contribuições positivas para a FCO_2 . Esse resultado é coerente, pois o pH regula a disponibilidade de nutrientes e a atividade enzimática microbiana, influenciando diretamente os processos de decomposição da matéria orgânica e a respiração heterotrófica do solo. A umidade relativa do ar (*umed*) e a pressão de vapor do solo (*pk_pa*) também apareceram entre as variáveis de maior impacto, enquanto o magnésio (*mg*) apresentou um ponto de alta influência positiva isolado, sugerindo efeito não linear e localizado dessa variável sobre determinadas observações. De modo geral, o *SVM* demonstrou maior sensibilidade a atributos químicos do solo em comparação aos demais modelos, provavelmente em razão da natureza de seu *kernel* de base radial, que captura relações de proximidade no espaço de características de forma diferente das estruturas de árvore.

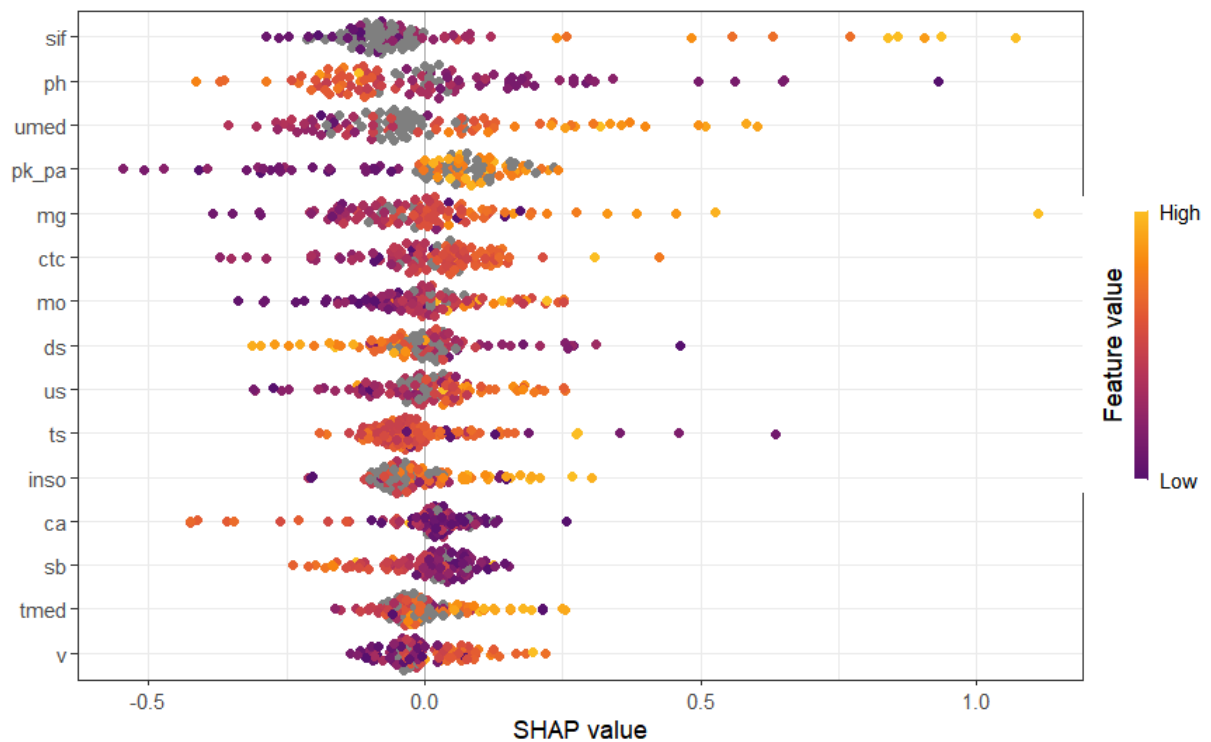


Figura 14 – Gráfico *beeswarm* dos valores SHAP para o modelo SVM.

O modelo *DT* (Figura 15) apresentou o padrão de valores SHAP mais concentrado em torno de zero entre todos os modelos avaliados, reflexo direto de sua estrutura que cria regiões de decisão abruptas e discretas, resultando em menor variação contínua das contribuições individuais. O potássio (k) foi identificado como a variável de maior importância pelo modelo, seguido pela SIF e pela velocidade mínima do vento (velmin). A relevância do potássio neste modelo pode estar associada ao seu papel na regulação do potencial osmótico das plantas e na atividade microbiana do solo, processos que influenciam indiretamente a respiração radicular e heterotrófica. A presença de velmin entre os principais preditores sugere que o *DT* capturou relações relacionadas à dinâmica atmosférica local, que pode afetar os gradientes de difusão de CO₂ entre o solo e a atmosfera. A menor amplitude dos valores SHAP e a maior proporção de observações com contribuição próxima de zero refletem a limitação do *DT* em capturar relações contínuas e complexas entre variáveis, resultado coerente com o menor desempenho preditivo observado para esse modelo ($R^2 = 0,61$).

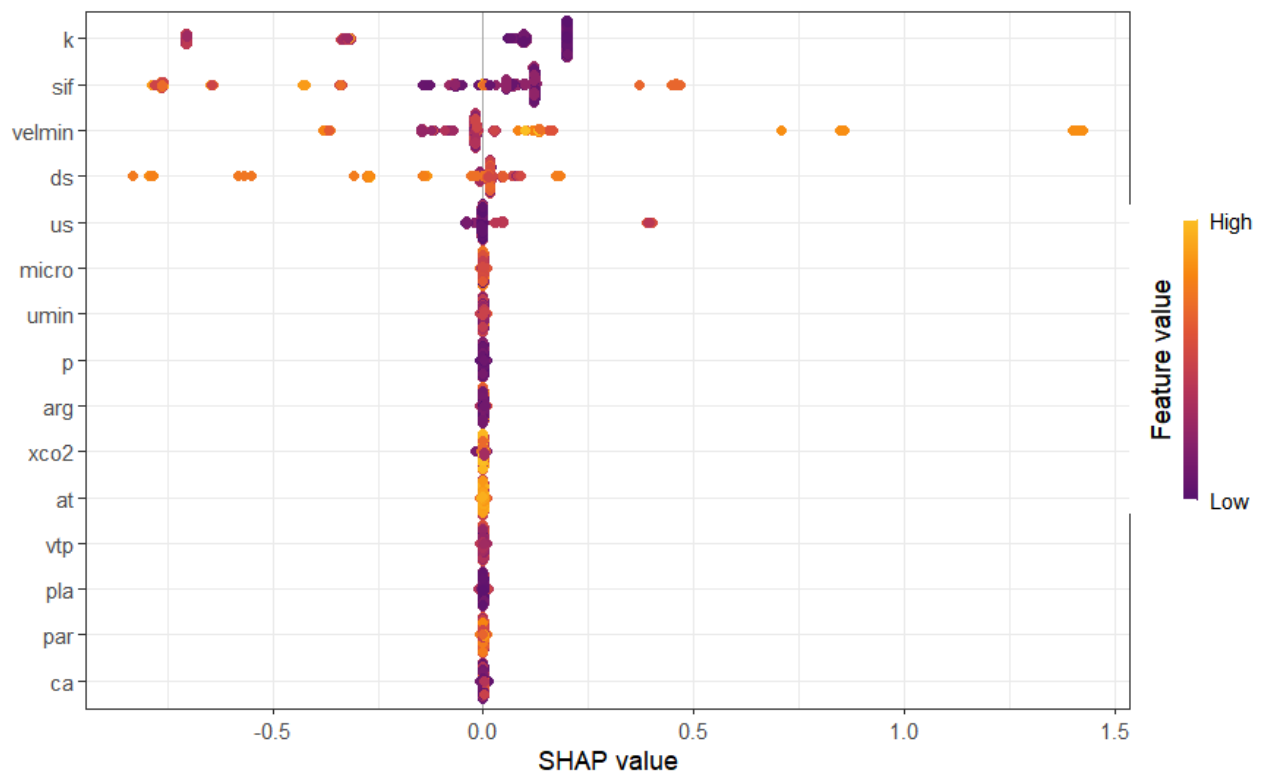


Figura 15 – Gráfico *beeswarm* dos valores SHAP para o modelo *DT*.

No modelo *RNA* (Figura 16), a densidade do solo (*ds*) foi identificada como a variável de maior importância, seguida pela *SIF* e pela velocidade mínima do vento (*velmin*). O destaque da densidade do solo neste modelo é notável, pois essa variável influencia diretamente a porosidade e a difusão de gases no perfil do solo, afetando tanto o transporte de CO_2 para a superfície quanto as condições aeróbicas para a atividade microbiana. A *RNA* apresentou uma distribuição de valores SHAP com menor amplitude geral em comparação ao *XGB* e ao *RF*, com predominância de contribuições positivas e relativamente simétricas. Variáveis como a pressão parcial do ar (*pk_pa*), o potássio (*k*) e o pH apareceram com importância relevante, indicando que a *RNA* capturou uma combinação distinta de atributos físico-químicos do solo e condições atmosféricas. A arquitetura de múltiplas camadas ocultas do modelo *multilayer perceptron* (MLP) permite aprender representações internas complexas das variáveis de entrada, o que pode explicar o padrão diferenciado de importância em relação aos modelos baseados em árvores.

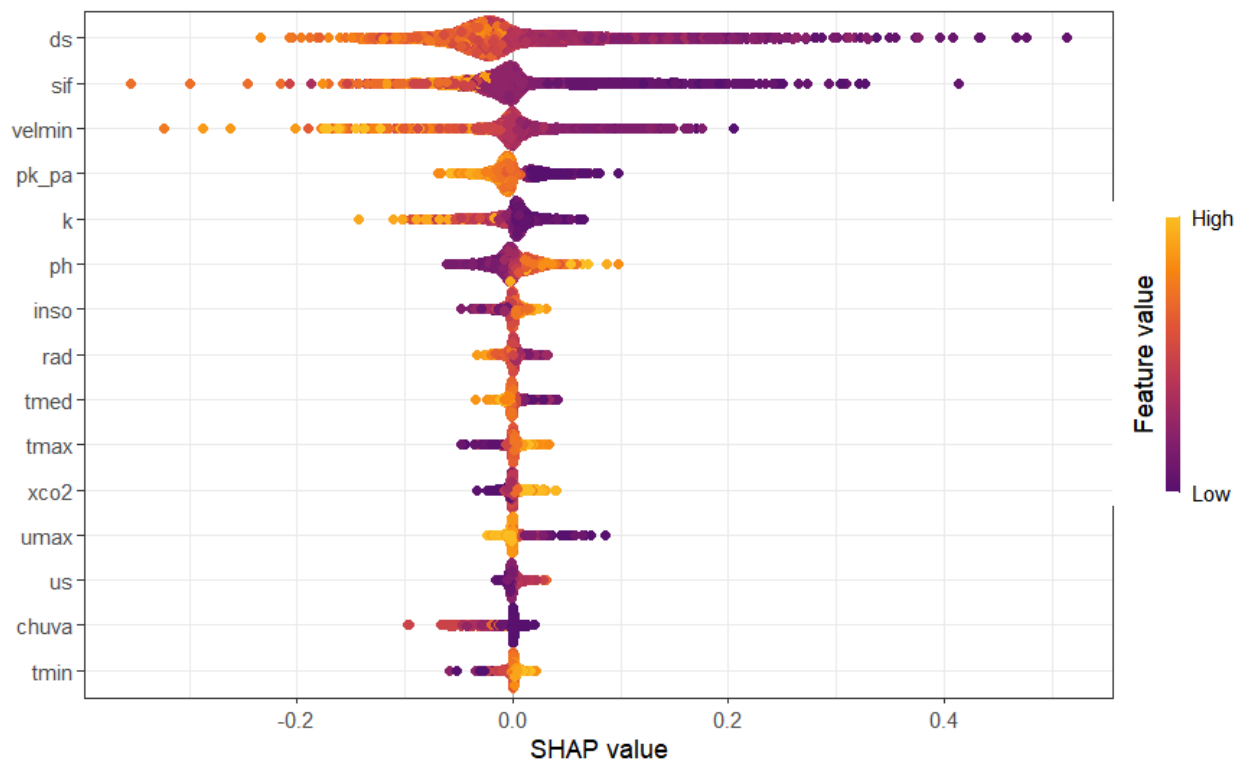


Figura 16 – Gráfico *beeswarm* dos valores SHAP para o modelo RNA.

De forma geral, a análise SHAP revelou que, apesar das diferenças estruturais entre os modelos, existe um conjunto consistente de variáveis que exerce grande influência sobre as previsões de FCO₂ em todos ou na maioria dos algoritmos avaliados. A SIF destacou-se como a variável mais recorrente e de maior impacto nos modelos *XGB*, *RF* e *SVM*, evidenciando a forte conexão entre a atividade fotossintética da vegetação e a respiração do solo. A xCO₂ e a umidade relativa do ar (umed) também apresentaram alta recorrência, indicando que as condições climáticas de umidade e a concentração atmosférica de CO₂ também foram importantes para determinar a dinâmica de emissão de CO₂ em áreas agrícolas do Cerrado.

5 CONSIDERAÇÕES FINAIS

Os modelos de aprendizado de máquina avaliados neste estudo demonstraram diferentes níveis de desempenho na predição dos fluxos de CO₂ do solo, refletindo tanto as particularidades de cada algoritmo quanto a complexidade dos processos bioquímicos envolvidos. Dentre os modelos avaliados, o *Random Forest* (RF) destacou-se como o mais eficiente, apresentando melhor desempenho preditivo em comparação aos demais algoritmos. Sua superioridade, ainda que sutil em relação a outros modelos, como o *XGBoost* (XGB), reforça sua robustez na modelagem de dados ambientais, especialmente por sua capacidade de capturar interações não lineares e lidar com alta variabilidade espacial e temporal. O modelo *XGBoost* (XGB) também apresentou desempenho elevado, confirmando sua eficiência como método baseado em árvores para modelagem de sistemas complexos. No entanto, seu maior custo computacional e a necessidade de ajuste de um número mais elevado de hiperparâmetros tornam sua implementação mais exigente quando comparada ao RF. As Redes Neurais Artificiais (RNA) e o *Support Vector Machine* (SVM) apresentaram desempenho intermediário, demonstrando boa capacidade de modelar relações não lineares, porém com maior sensibilidade à variabilidade dos dados e menor robustez em relação aos modelos *ensemble*. Já o modelo de *Decision Tree* (DT) apresentou o desempenho mais limitado, o que era esperado devido à sua estrutura simplificada e maior suscetibilidade ao sobreajuste. Em relação à importância das variáveis, observou-se que indicadores da atividade da vegetação (como a SIF), variáveis climáticas (temperatura e umidade) e atributos químicos do solo estiveram entre os principais preditores em praticamente todos os modelos. Em especial, fluorescência de clorofila induzida pelo sol (SIF) destacou-se como uma das variáveis mais importantes na maioria dos modelos, evidenciando sua forte associação com a dinâmica do carbono no sistema solo–planta–atmosfera. Considerando que a SIF está diretamente relacionada à atividade fotossintética da vegetação e ao controle do CO₂ atmosférico, sua elevada relevância sugere uma forte conexão entre os processos de assimilação de carbono e a emissão de CO₂ pelo solo. Além disso, variáveis climáticas, como temperatura e umidade, também apresentaram elevada importância, reforçando seu papel no controle da atividade microbiana, da respiração radicular e da difusão de gases no solo. Os resultados sugerem que os fluxos de CO₂ do solo estão mais associados à respiração autotrófica, ligada à atividade das raízes e das plantas, do que à respiração microbiana. A alta importância de variáveis como a SIF

reforça essa ideia e indica que a dinâmica do carbono no solo está diretamente ligada à fotossíntese e à entrada recente de carbono no sistema.

6 REFERÊNCIAS

ALMEIDA, R. F. D.; TEIXEIRA, D. D. B.; MONTANARI, R.; BOLONHEZI, A. C.; TEIXEIRA, E. B.; MOITINHO, M. R.; PANOSSO, A. R.; SPOKAS, K. A.; JR, N. L. S. Ratio of CO₂ and O₂ as index for categorising soil biological activity in sugarcane areas under contrasting straw management regimes. **Soil Research**, v. 56, n. 4, p. 373-381, 2018.

APPEEARS TEAM. **Application for Extracting and Exploring Analysis Ready Samples (AppEEARS)**. Ver. 6. NASA EOSDIS Land Processes Distributed Active Archive Center (LP DAAC), USGS/Earth Resources Observation and Science (EROS) Center, Sioux Falls, South Dakota, USA, 2020.

BICALHO, E. S.; MOITINHO, M. R.; DE BORTOLI TEIXEIRA, D.; PANOSSO, A. R.; SPOKAS, K. A.; LA SCALA JR, N. Soil greenhouse gases: relations to soil attributes in a sugarcane production area. **Soil Science Society of America Journal**, v. 81, n. 5, p. 1168-1178, 2017.

BOATENG, E. B.; PILLAY, M.; DAVIS, P. Predicting the level of safety performance using an artificial neural network. In: AHRAM, T.; KARWOWSKI, W.; TAIAR, R. (Eds.). **Human Systems Engineering and Design**. Cham: Springer International Publishing, 2019. p. 705-710.

BREIMAN, L.; CUTLER, A.; LIAW, A.; WIENER, M. **randomForest: Breiman and Cutler's Random Forests for Classification and Regression**. The Comprehensive R Archive Network, 2022.

CAMPOS, M. C. C.; ALHO, L. C.; SILVA, D. A. P. D.; SOARES, M. D. R.; CUNHA, J. M. D.; SILVA, D. M. P. D. Distribuição espacial do efluxo de CO₂ em área de terra preta arqueológica sob cultivo de cacau e café no município de Apuí, AM, Brasil. **Revista Ambiente & Água**, v. 11, n., p. 788-798, 2016.

CANTERL, K. F. F.; VICENTINI, M. E.; DE LUCENA, W. B.; DE MORAES, M. L. T.; MONTANARI, R.; FERRAUDO, A. S.; PERUZZI, N. J.; LA SCALA, N.; PANOSSO, A. R. Machine learning for prediction of soil CO₂ emission in tropical forests in the Brazilian Cerrado. **Environmental Science and Pollution Research**, v., n., p., 2023.

CERRI, C. E. P.; SPAROVEK, G.; BERNOUX, M.; EASTERLING, W. E.; MELILLO, J. M.; CERRI, C. C. Tropical agriculture and global warming: impacts and mitigation options. **Scientia Agricola**, v. 64, n. 1, p. 83-99, 2007.

CHAPMAN, P.; CLINTON, J.; KERBER, R.; KHABAZA, T.; REINARTZ, T.; SHEARER, C. R.; WIRTH, R. **CRISP-DM 1.0: Step-by-step data mining guide**. 2000.

CHEN, T.; HE, T.; BENESTY, M.; KHOTILOVICH, V.; TANG, Y.; CHO, H.; CHEN, K.; MITCHELL, R.; CANO, I.; ZHOU, T.; LI, M.; XIE, J.; LIN, M.; GENG, Y.; LI, Y.; YUAN, J. **xgboost: Extreme Gradient Boosting**. The Comprehensive R Archive Network, 2023.

COSTA, L. M. D.; SANTOS, G. A. D. A.; PANOSSO, A. R.; ROLIM, G. D. S.; SCALA, N. L. Daily model of XCO₂ based on orbital observations for the State of São Paulo – Brazil. **Research Square**, 2022.

CRISP, D.; FISHER, B. M.; O'DELL, C.; FRANKENBERG, C.; BASILIO, R.; BÖSCH, H.; BROWN, L. R.; CASTANO, R.; CONNOR, B.; DEUTSCHER, N. M.; ELDERING, A.; GRIFFITH, D.; GUNSON, M.; KUZE, A.; MANDRAKE, L.; MCDUFFIE, J.; MESSERSCHMIDT, J.; MILLER, C. E.; MORINO, I.; NATRAJ, V.; NOTHOLT, J.; O'BRIEN, D. M.; OYAFUSO, F.; POLONSKY, I.; ROBINSON, J.; SALAWITCH, R.; SHERLOCK, V.; SMYTH, M.; SUTO, H.; TAYLOR, T. E.; THOMPSON, D. R.; WENNERBERG, P. O.; WUNCH, D.; YUNG, Y. L. The ACOS CO₂ retrieval algorithm – Part II: global XCO₂ data characterization. **Atmos. Meas. Tech.**, v. 5, n. 4, p. 687-707, 2012.

DA COSTA, L. M.; DE ARAÚJO SANTOS, G. A.; DE MENDONÇA, G. C.; MORAIS FILHO, L. F. F.; DE MENESES, K. C.; DE SOUZA ROLIM, G.; LA SCALA JR, N. Spatiotemporal variability of atmospheric CO₂ concentration and controlling factors over sugarcane cultivation areas in southern Brazil. **Environment, Development and Sustainability**, v., n., p., 2021.

DA COSTA, L. M.; SANTOS, G. A. D.; PANOSSO, A. R.; ROLIM, G. D.; LA SCALA, N. An empirical model for estimating daily atmospheric column-averaged CO₂ concentration above Sao Paulo state, Brazil. **Carbon Balance and Management**, v. 17, n. 1, p., 2022.

DA COSTA, L. M.; DE MENDONÇA, G. C.; ARAÚJO SANTOS, G. A. D.; MORAES, J. R. D. S. C. D.; COLOMBO, R.; PANOSSO, A. R.; LA SCALA JR, N. High spatial resolution solar-induced chlorophyll fluorescence and its relation to rainfall precipitation across Brazilian ecosystems. **Environmental Research**, v. 218, n., p. 114991, 2023.

DA CUNHA, J. M.; CAMPOS, M. C. C.; GAIO, D. C.; DE SOUZA, Z. M.; SOARES, M. D. R.; DA SILVA, D. M. P.; SIMÕES, E. L. Spatial variability of soil respiration in Archaeological Dark Earth areas in the Amazon. **Catena**, v. 162, n., p. 148-156, 2018.

DE FIGUEIREDO, E. B.; JAYASUNDARA, S.; DE OLIVEIRA BORDONAL, R.; BERCHIELLI, T. T.; REIS, R. A.; WAGNER-RIDDLE, C.; LA SCALA JR, N. Greenhouse gas balance and carbon footprint of beef cattle in three contrasting pasture-management systems in Brazil. **Journal of Cleaner Production**, v. 142, n., p. 420-431, 2017.

DE LUCENA, W. B.; VICENTINI, M. E.; DE ARAÚJO SANTOS, G. A.; DE OLIVEIRA SILVA, B.; MESQUITA DA COSTA, D. V.; FERREIRA CANTERL, K. F.; NEIRA ROMÁN, J. A.; DE SOUZA ROLIM, G.; PANOSSO, A. R.; LA SCALA, N. Temporal variability of the CO₂ emission and the O₂ influx in a tropical soil in contrasting coverage conditions. **Journal of South American Earth Sciences**, v. 121, n. 1, p. 104120, 2023.

DE SOUZA MARIA, L.; ROSSI, F. S.; COSTA, L. M. D.; CAMPOS, M. O.; BLAS, J. C. G.; PANOSSO, A. R.; SILVA, J. L. D.; SILVA JUNIOR, C. A. D.; LA SCALA JR, N. Spatiotemporal analysis of atmospheric XCH₄ as related to fires in the Amazon biome

during 2015–2020. **Remote Sensing Applications: Society and Environment**, v. 30, n., p. 100967, 2023.

EMBRAPA. **Manual de métodos de análise de solo**. 2. ed. Brasília: Ministério da Agricultura e do Abastecimento/EMBRAPA-CNPq, 1997. 212 p.

FARHATE, C. V. V.; SOUZA, Z. M. D.; OLIVEIRA, S. R. D. M.; CARVALHO, J. L. N.; SCALA JÚNIOR, N. L.; SANTOS, A. P. G. Classification of soil respiration in areas of sugarcane renewal using decision tree. **Scientia Agricola**, v. 75, n., p. 216-224, 2018.

FREITAS, L. P. S.; LOPES, M. L. M.; CARVALHO, L. B.; PANOSSO, A. R.; LA SCALA JÚNIOR, N.; FREITAS, R. L. B.; MINUSSI, C. R.; LOTUFO, A. D. P. Forecasting the spatiotemporal variability of soil CO₂ emissions in sugarcane areas in southeastern Brazil using artificial neural networks. **Environmental Monitoring and Assessment**, v. 190, n. 12, p. 741, 2018.

GHIASI, M. M.; ARABLOO, M.; MOHAMMADI, A. H.; BARGHI, T. Application of ANFIS soft computing technique in modeling the CO₂ capture with MEA, DEA, and TEA aqueous solutions. **International Journal of Greenhouse Gas Control**, v. 49, n., p. 47-54, 2016.

GRUNWALD, S. Artificial intelligence and soil carbon modeling demystified: power, potentials, and perils. **Soil Security**, v. 1, n. 1, p. 5, 2022.

GUARENGHI, M. M.; GAROFALO, D. F. T.; SEABRA, J. E. A.; MOREIRA, M. M. R.; NOVAES, R. M. L.; RAMOS, N. P.; NOGUEIRA, S. F.; DE ANDRADE, C. A. Land use change net removals associated with sugarcane in Brazil. **Land**, v. 12, n. 3, p. 584, 2023.

HAMRANI, A.; AKBARZADEH, A.; MADRAMOOTOO, C. A. Machine learning for predicting greenhouse gas emissions from agricultural soils. **Science of the Total Environment**, v. 741, n., p. 140338, 2020.

HARSÁNYI, E.; MIRZAEI, M.; ARSHAD, S.; ALSILIBE, F.; VAD, A.; NAGY, A.; RATONYI, T.; GORJI, M.; AL-DALAHME, M.; MOHAMMED, S. Assessment of advanced machine and deep learning approaches for predicting CO₂ emissions from agricultural lands: insights across diverse agroclimatic zones. **Earth Systems and Environment**, v., n., p., 2024.

HAYKIN, S. **Redes neurais: princípios e prática**. Porto Alegre: Bookman, 2001.

IPCC. **Climate Change 2013: The Physical Science Basis**. New York: Cambridge University Press, 2013. 1535 p.

JAMES, G.; WITTEN, D.; HASTIE, T.; TIBSHIRANI, R. **An Introduction to Statistical Learning – with Applications in R**. New York: Springer, 2013.

JANG, J. S. R. ANFIS: adaptive-network-based fuzzy inference system. **IEEE Transactions on Systems, Man, and Cybernetics**, v. 23, n. 3, p. 665-685, 1993.

KROESE, D. P.; BOTEV, Z. I.; TAIMRE, T.; VAISMAN, R. **Data Science and Machine Learning: Mathematical and Statistical Methods**. Boca Raton: Chapman and Hall/CRC, 2019.

KUHN, M.; WICKHAM, H. **Tidymodels: a collection of packages for modeling and machine learning using tidyverse principles**. The Comprehensive R Archive Network, 2020.

LUNDBERG, S. M.; LEE, S. I. A unified approach to interpreting model predictions. **Advances in Neural Information Processing Systems**, v. 30, p. 4765-4774, 2017.

LUNDBERG, S. M. et al. From local explanations to global understanding with explainable AI for trees. **Nature Machine Intelligence**, v. 2, n. 1, p. 56-67, 2020.

MANTOVANELLI, B. C.; CAMPOS, M. C. C.; ALHO, L. C.; SILVA, P. C.; SILVA, D. A. P.; CUNHA, J. M.; SILVA, D. M. P.; SOARES, M. D. R. Distribuição espacial da emissão de CO₂ e atributos do solo sob campo nativo na região de Humaitá, Amazonas. **Sociedade & Natureza**, v. 28, n. 2, p., 2016.

MAPBIOMAS. **MapBiomas Project – Collection 6.0 of the Annual Series of Land Use and Land Cover Maps of Brazil**. 2021.

MASSIE, S. T.; CRONK, H.; MERRELLI, A.; O'DELL, C.; SCHMIDT, K. S.; CHEN, H.; BAKER, D. Analysis of 3D cloud effects in OCO-2 XCO₂ retrievals. **Atmos. Meas. Tech.**, v. 14, n. 2, p. 1475-1499, 2021.

MCBRATNEY, A.; DE GRUIJTER, J.; BRYCE, A. Pedometrics timeline. **Geoderma**, v. 338, n., p. 568-575, 2019.

MCBRATNEY, A. B.; SANTOS, M. L. M.; MINASNY, B. On digital soil mapping. **Geoderma**, v. 117, n. 1-2, p. 3-52, 2003.

MOHAMMED, G. H.; COLOMBO, R.; MIDDLETON, E. M.; RASCHER, U.; VAN DER TOL, C.; NEDBAL, L.; GOULAS, Y.; PÉREZ-PRIEGO, O.; DAMM, A.; MERONI, M.; JOINER, J.; COGLIATI, S.; VERHOEF, W.; MALENOSKÝ, Z.; GASTELLU-ETCHEGORRY, J.-P.; MILLER, J. R.; GUANTER, L.; MORENO, J.; MOYA, I.; BERRY, J. A.; FRANKENBERG, C.; ZARCO-TEJADA, P. J. Remote sensing of solar-induced chlorophyll fluorescence (SIF) in vegetation: 50 years of progress. **Remote Sensing of Environment**, v. 231, n., p. 111177, 2019.

MOITINHO, M. R.; FERRAUDO, A. S.; PANOSSO, A. R.; BICALHO, E. D. S.; TEIXEIRA, D. D. B.; BARBOSA, M. D. A.; TSAI, S. M.; BORGES, B. M. F.; CANNAVAN, F. D. S.; SOUZA, J. A. M. D.; LA SCALA, N. Effects of burned and unburned sugarcane harvesting systems on soil CO₂ emission and soil physical, chemical, and microbiological attributes. **Catena**, v. 196, n., p. 104903, 2021a.

MOITINHO, M. R.; FERRAUDO, A. S.; PANOSSO, A. R.; BICALHO, E. D. S.; TEIXEIRA, D. D. B.; BARBOSA, M. D. A.; TSAI, S. M.; FERRARI BORGES, B. M.; CANNAVAN, F. D. S.; MARCONDES DE SOUZA, J. A.; LA SCALA JR, N. Effects of

burned and unburned sugarcane harvesting systems on soil CO₂ emission and soil physical, chemical, and microbiological attributes. **Catena**, v. 196, n., p., 2021b.

MOITINHO, M. R.; TEIXEIRA, D. D. B.; BICALHO, E. D. S.; PANOSSO, A. R.; FERRAUDO, A. S.; PEREIRA, G. T.; TSAI, S. M.; BORGES, B. M. F.; LA SCALA JR, N. Soil CO₂ emission and soil attributes associated with the microbiota of a sugarcane area in southern Brazil. **Scientific Reports**, v. 11, n. 1, p. 8325, 2021c.

MORAIS FILHO, L. F. F.; MENESES, K. C. D.; SANTOS, G. A. D. A.; BICALHO, E. D. S.; ROLIM, G. D. S.; LA SCALA JR, N. xCO₂ temporal variability above Brazilian agroecosystems: a remote sensing approach. **Journal of Environmental Management**, v. 288, n., p. 112433, 2021.

NIKITENKO, A. A.; TIMOFEEV, Y. M.; BEREZIN, I. A.; POBEROVSKII, A. V.; VIROLAINEN, Y. A.; POLYAKOV, A. V. The analysis of OCO-2 satellite measurements of CO₂ in the vicinity of Russian cities. **Atmospheric and Oceanic Optics**, v. 33, n. 6, p. 650-655, 2020.

O'DELL, C. W.; CONNOR, B.; BÖSCH, H.; O'BRIEN, D.; FRANKENBERG, C.; CASTANO, R.; CHRISTI, M.; ELDERING, D.; FISHER, B.; GUNSON, M.; MCDUFFIE, J.; MILLER, C. E.; NATRAJ, V.; OYAFUSO, F.; POLONSKY, I.; SMYTH, M.; TAYLOR, T.; TOON, G. C.; WENNERBERG, P. O.; WUNCH, D. The ACOS CO₂ retrieval algorithm – Part 1: description and validation against synthetic observations. **Atmos. Meas. Tech.**, v. 5, n. 1, p. 99-121, 2012.

OLIVEIRA, P. H. G. **Modelagem de equações estruturais aplicada ao estudo das relações entre qualidade e diversidade da fauna no sistema solo-serapilheira**. 2016. Tese (Doutorado) – Universidade Estadual do Norte Fluminense Darcy Ribeiro, Campos dos Goytacazes, 2016.

PANOSSO, A. R. **Emissão de CO₂ do solo em áreas agrícolas: abordagem em aprendizado de máquina estatístico**. 2023. 92 p.

PAUSTIAN, K.; LEHMANN, J.; OGLE, S.; REAY, D.; ROBERTSON, G. P.; SMITH, P. Climate-smart soils. **Nature**, v. 532, n. 7597, p. 49-57, 2016.

PINHEIRO DA SILVA, D. A.; CAMPOS, M. C. C.; MANTOVANELLI, B. C.; SANTOS, L. A. C. D.; SOARES, M. D. R.; CUNHA, J. M. D. Variabilidade espacial da emissão de CO₂, temperatura e umidade do solo em área de pastagem na região Amazônica, Brasil. **Revista de Ciências Agroveterinárias**, v. 18, n. 1, p. 119-126, 2019.

POSIT TEAM. **RStudio: Integrated Development Environment for R**. Boston: Posit Software, PBC. Disponível em: <http://www.posit.co/>. 2023.

R DEVELOPMENT CORE TEAM. **A Language and Environment for Statistical Computing**. Vienna: R Foundation for Statistical Computing. ISBN 3-900051-07-0. Disponível em: <http://www.R-project.org/>. 2023.

RAIJ, B. V. **Análise química para avaliação da fertilidade de solos tropicais**. Campinas: Instituto Agronômico, 2001. 285 p.

REICHSTEIN, M.; CAMPS-VALLS, G.; STEVENS, B.; JUNG, M.; DENZLER, J.; CARVALHAIS, N.; PRABHAT. Deep learning and process understanding for data-driven Earth system science. **Nature**, v. 566, n. 7743, p. 195-204, 2019.

ROSSI, F. S.; DE ARAÚJO SANTOS, G. A.; DE SOUZA MARIA, L.; LOURENÇONI, T.; PELISSARI, T. D.; DELLA-SILVA, J. L.; OLIVEIRA JÚNIOR, J. W.; SILVA, A. D. A. E.; LIMA, M.; TEODORO, P. E.; TEODORO, L. P. R.; DE OLIVEIRA-JÚNIOR, J. F.; LA SCALA JR, N.; SILVA JUNIOR, C. A. D. Carbon dioxide spatial variability and dynamics for contrasting land uses in central Brazil agricultural frontier from remote sensing data. **Journal of South American Earth Sciences**, v. 116, n., p. 103809, 2022.

SEEG. **Emissões de GEE do Brasil e suas implicações para políticas públicas e a contribuição brasileira para o Acordo de Paris**. Observatório do Clima, 2018. 80 p.

SHAPLEY, L. S. A value for n-person games. In: KUHN, H. W.; TUCKER, A. W. (Eds.). **Contributions to the Theory of Games**. Princeton: Princeton University Press, 1953. v. 2, p. 307-317.

SILVA, B. D.; MOITINHO, M. R.; SANTOS, G. A. D.; TEIXEIRA, D. D.; FERNANDES, C.; LA SCALA, N. Soil CO₂ emission and short-term soil pore class distribution after tillage operations. **Soil & Tillage Research**, v. 186, n., p. 224-232, 2019.

SOUZA, C. M.; SHIMBO, J. Z.; ROSA, M. R.; PARENTE, L. L.; ALENCAR, A. A.; RUDORFF, B. F. T.; HASENACK, H.; MATSUMOTO, M.; FERREIRA, L. G.; SOUZA-FILHO, P. W. M.; DE OLIVEIRA, S. W.; ROCHA, W. F.; FONSECA, A. V.; MARQUES, C. B.; DINIZ, C. G.; COSTA, D.; MONTEIRO, D.; ROSA, E. R.; VÉLEZ-MARTIN, E.; WEBER, E. J.; LENTI, F. E. B.; PATERNOST, F. F.; PAREYN, F. G. C.; SIQUEIRA, J. V.; VIERA, J. L.; NETO, L. C. F.; SARAIVA, M. M.; SALES, M. H.; SALGADO, M. P. G.; VASCONCELOS, R.; GALANO, S.; MESQUITA, V. V.; AZEVEDO, T. Reconstructing three decades of land use and land cover changes in Brazilian biomes with Landsat Archive and Earth Engine. **Remote Sensing**, v. 12, n. 17, p. 2735, 2020.

SPECHT, D. F. A general regression neural network. **IEEE Transactions on Neural Networks**, v. 2, n. 6, p. 568-576, 1991.

STACKHOUSE, P. W. J.; WESTBERG, D.; CHANDLER, W. S.; ZHANG, T.; HOELL, J. M. **Prediction of Worldwide Energy Resource (POWER) Agroclimatology Methodology Version 1.1.0**. NASA Langley Research Center, 2015.

TAVANTI, R. F. R.; MONTANARI, R.; PANOSSO, A. R.; LA SCALA, N.; CHIQUITELLI NETO, M.; FREDDI, O. D. S.; PAZ GONZÁLEZ, A.; DE CARVALHO, M. A. C.; SOARES, M. B.; TAVANTI, T. R.; GALINDO, F. S. What is the impact of pasture reform on organic carbon compartments and CO₂ emissions in the Brazilian Cerrado? **Catena**, v. 194, n., p. 104702, 2020.

TAVARES, R. L. M.; OLIVEIRA, S. R. D.; DE BARROS, F. M. M.; FARHATE, C. V. V.; DE SOUZA, Z. M.; LA SCALA, N. Prediction of soil CO₂ flux in sugarcane management

systems using the Random Forest approach. **Scientia Agricola**, v. 75, n. 4, p. 281-287, 2018.

TAVARES, R. L. M.; SOUZA, Z. M.; SIQUEIRA, D. S.; LA SCALA JR, N.; PANOSSO, A. R.; CAMPOS, M. C. C. Soil CO₂ emission in sugarcane management systems. **Acta Agriculturae Scandinavica, Section B – Soil & Plant Science**, v. 1, n. 1, p. 8, 2015.

THERNEAU, T.; ATKINSON, B.; RIPLEY, B. **rpart: Recursive Partitioning and Regression Trees**. The Comprehensive R Archive Network, 2022.

USSIRI, D. A. N.; LAL, R. Long-term tillage effects on soil carbon storage and carbon dioxide emissions in continuous corn cropping system from an alfisol in Ohio. **Soil & Tillage Research**, v. 104, n. 1, p. 39-47, 2009.

VICENTINI, M. E.; PINOTTI, C. R.; HIRAI, W. Y.; DE MORAES, M. L. T.; MONTANARI, R.; FILHO, M. C. M. T.; MILORI, D. M. B. P.; JÚNIOR, N. L. S.; PANOSSO, A. R. CO₂ emission and its relation to soil temperature, moisture, and O₂ absorption in the reforested areas of Cerrado biome, Central Brazil. **Plant and Soil**, v., n., p., 2019.

WICKHAM, H.; AVERICK, M.; BRYAN, J.; CHANG, W.; MCGOWAN, L.; FRANÇOIS, R.; GROLEMUND, G.; HAYES, A.; HENRY, L.; HESTER, J.; KUHN, M.; PEDERSEN, T.; MILLER, E.; BACHE, S.; MÜLLER, K.; OOMS, J.; ROBINSON, D.; SEIDEL, D.; SPINU, V.; TAKAHASHI, K.; VAUGHAN, D.; WILKE, C.; WOO, K.; YUTANI, H. Welcome to the tidyverse. **Journal of Open Source Software**, v. 4, n. 43, p. 1686, 2019.

WMO. **WMO Greenhouse Gas Bulletin (GHG Bulletin) – No. 19: The State of Greenhouse Gases in the Atmosphere Based on Global Observations through 2022**. Tokyo: World Meteorological Organization, World Data Centre for Greenhouse Gases. Disponível em: https://library.wmo.int/viewer/68532/download?file=GHG-19_en.pdf. 2023.

XAVIER, C. V.; MOITINHO, M. R.; TEIXEIRA, D. D. B.; DE ARAÚJO SANTOS, G. A.; CORÁ, J. E.; LA SCALA JR, N. Crop rotation and sequence effects on temporal variation of CO₂ emissions after long-term no-till application. **Science of the Total Environment**, v. 709, n., p. 136107, 2020.

XIONG, X.; GRUNWALD, S.; MYERS, D. B.; KIM, J.; HARRIS, W. G.; COMERFORD, N. B. Holistic environmental soil-landscape modeling of soil organic carbon. **Environmental Modelling & Software**, v. 57, n., p. 202-215, 2014.

YILMAZ, G.; BILGILI, A. V. Modeling seasonal variations of long-term soil CO₂ emissions in an orchard plantation in a semiarid area, SE Turkey. **Environmental Monitoring and Assessment**, v. 190, n. 8, p., 2018.

YILMAZ, I.; KAYNAR, O. Multiple regression, ANN (RBF, MLP) and ANFIS models for prediction of swell potential of clayey soils. **Expert Systems with Applications**, v. 38, n. 5, p. 5958-5966, 2011.