Universidade de Brasília
Faculdade de Tecnologia
Departamento de Engenharia Mecânica

Predição de emissões de óxido nitroso em solos cultivados com cana-de-açúcar utilizando modelos de aprendizado de máquina.

Rafael Teixeira Bonato

DISSERTAÇÃO DE MESTRADO

PROGRAMA DE PÓS-GRADUAÇÃO EM SISTEMAS MECATRÔNICOS

Brasília

2026

Universidade de Brasília
Faculdade de Tecnologia
Departamento de Engenharia Mecânica

Prediction of nitrous oxide emissions in soils cultivated with sugarcane using machine learning models.

Rafael Teixeira Bonato

Dissertação de Mestrado submetida ao Departamento de Engenharia Mecânica da Universidade Brasília como parte dos requisitos necessários para a obtenção do grau de Mestre

Orientador: Prof. Dr. Díbio L.Borges

Brasília
2026

Teixeira Bonato , Rafael.
T769p Predição de emissões de óxido nitroso em solos cultivados com cana-de-açúcar utilizando modelos de aprendizado de máquina. / Rafael Teixeira Bonato ; orientador Díbio L.Borges. -- Brasília, 2026.
50 p.

Dissertação de Mestrado (Programa de Pós-Graduação em Sistemas Mecatrônicos) -- Universidade de Brasília, 2026.

1. Aprendizado de máquina. 2. GEE. 3. Cana-de-açúcar. 4. Irrigação. I. Díbio L.Borges, , orient. II. Título

Universidade de Brasília
Faculdade de Tecnologia
Departamento de Engenharia Mecânica

Predição de emissões de óxido nitroso em solos cultivados com cana-de-açúcar utilizando modelos de aprendizado de máquina.

Rafael Teixeira Bonato

Dissertação de Mestrado submetida ao Departamento de Engenharia Mecânica da Universidade Brasília como parte dos requisitos necessários para a obtenção do grau de Mestre

Trabalho aprovado. Brasília, 20 de janeiro de 2026:

Prof. Dr. Díbio Leandro Borges, UnB/FT/CIC
Presidente/Orientador

Prof. Dr. José Maurício Santos Torres da Motta, UnB/FT/ENM
Examinador interno

Prof. Dra. Arminda Moreira de Carvalho, Embrapa
Examinador externo

Brasília
2026

Dedico este trabalho aos meus pais, Maria Cristina e Paulo, e à minha irmã Isabella, cujos valores e incentivo me trouxeram até aqui. Ao João, pela presença constante, paciência e por caminhar ao meu lado na construção deste sonho.

Agradecimentos

Agradecemos ao Dr. Díbio Leandro Borges pela orientação criteriosa, constante disponibilidade e pelas valiosas contribuições científicas ao longo de todas as etapas desta pesquisa.

Agradecemos também à Profa. Dra. Arminda Moreira de Carvalho e à equipe da Embrapa pela cessão dos dados utilizados neste estudo, bem como pelo apoio institucional, sem os quais a realização desta pesquisa não seria possível. As opiniões e análises fornecidas pela equipe da Embrapa foram essenciais para o adequado enquadramento dos resultados no contexto agronômico, fortalecendo a consistência científica e a aplicabilidade prática deste trabalho.

Resumo

O óxido nitroso (N_2O) é um relevante gás de efeito estufa, com cerca de 60% de suas emissões associadas a atividades agrícolas. Seus fluxos no solo são influenciados por fatores como aplicação de fertilizantes nitrogenados, disponibilidade de nitrogênio, preparo do solo, temperatura, pH e umidade, que interagem de forma não linear. Modelos preditivos baseados em aprendizado de máquina podem contribuir para estimar essas emissões, melhorar a compreensão do fenômeno e apoiar estratégias de mitigação.

Este estudo aplicou técnicas de aprendizado de máquina para estimar fluxos de N_2O em cultivo de cana-de-açúcar sob diferentes regimes de irrigação, comparando os resultados com análises convencionais. O algoritmo *Random Forest* apresentou desempenho superior, com coeficiente de determinação (R^2) de 87,36%. Os resultados evidenciam a dificuldade de predição em cenários com pequenos volumes de dados, dada a distribuição da variável estudada. Os modelos utilizados apresentam-se como uma alternativa viável aos métodos clássicos, permitindo uma análise mais abrangente dos dados. Além disso, o comportamento observado alinha-se à literatura existente, o que corrobora os achados deste estudo.

Palavras-chave: Aprendizado de máquina. GEE. Cana-de-açúcar. Irrigação.

Abstract

Nitrous oxide (N_2O) is a significant greenhouse gas, with approximately 60% of its emissions associated with agricultural activities. Its soil fluxes are influenced by factors such as nitrogen fertilizer application, nitrogen availability, soil preparation, temperature, pH, and moisture, which interact in a non-linear manner. Machine learning-based predictive models can help estimate these emissions, improve the understanding of the phenomenon, and support mitigation strategies.

This study applied machine learning techniques to estimate N_2O fluxes in sugarcane cultivation under different irrigation regimes, comparing the results with conventional analyses. The *Random Forest* algorithm presented superior performance, with a coefficient of determination (R^2) of 87.36%. The results highlight the difficulty of prediction in scenarios with small data volumes, given the distribution of the studied variable. The models used present themselves as a viable alternative to classical methods, allowing for a more comprehensive analysis of the data. Furthermore, the observed behavior aligns with existing literature, corroborating the findings of this study.

Keywords: Machine learning. Nitrous oxide. Sugarcane. Irrigation.

Lista de ilustrações

Lista de tabelas

Sumário

1 Introdução

A temática sobre as emissões de gases do efeito estufa (GEE) que contribuem para o aquecimento global tem adquirido crescente destaque na literatura contemporânea, especialmente diante de catástrofes climáticas e impactos na segurança alimentar gerado pela mudança ambientais. Nesse cenário, compreender o impacto negativo do cultivo de cana-de-açúcar na liberação de substâncias nocivas ao meio ambiente torna-se fundamental para o avanço das pesquisas na área de gestão do manejo agrícola com foco na sustentabilidade e preservação ambiental. Sob essa ótica, este capítulo discute o efeito de gases emitidos pelo solo durante o processo de produção agrícola, os quais contribuem significativamente para o efeito estufa. Primeiramente, são apresentados estudos que buscam quantificar o volume de emissões em culturas de elevada relevância econômica no Brasil, como a cana-de-açúcar. Além disso, examina-se de que forma diferentes práticas de manejo agrícola podem influenciar no aumento ou na redução dessas emissões. No estudo em análise, foram avaliados distintos regimes de manejo hídrico, com o objetivo de compreender seus impactos tanto na produtividade quanto na liberação de N_2O.

1.1 Caracterização do Problema

O óxido nitroso (N_2O) é um gás de efeito estufa com potencial de aquecimento global aproximadamente 298 vezes superior ao do dióxido de carbono (CO_2). Ele é o terceiro gás mais relevante em termos de forçamento radiativo e o principal responsável pela destruição da camada de ozônio(Portmann; Daniel; Ravishankara, 2012). Aproximadamente 60% das emissões globais de N_2O têm origem em atividades humanas, destacando-se a agricultura, onde a aplicação de fertilizantes nitrogenados e as condições do solo e do clima desempenham papel determinante. Dessa forma, essas condições edafoclimáticas também sofrem interveniência do manejo da cultura, com a irrigação e movimentação do solo (Tian *et al.*, 2020).

Na agricultura, o N_2O é produzido predominantemente por processos microbianos de nitrificação e desnitrificação no solo, cuja intensidade tende a aumentar após a aplicação de fertilizantes nitrogenados (Butterbach-Bahl; Dannenmann, 2011). Fatores como disponibilidade de nitrogênio, temperatura, pH, umidade do solo, tipo de solo e manejo agrícola interagem de forma complexa, influenciando diretamente a magnitude das emissões. A mitigação da emissão de N_2O é fundamental para a redução do impacto ambiental, visto o potencial de aquecimento global (PAG) deste gás , tornando a sustentabilidade do sistema o foco principal, para além da retenção de nitrogênio (Yin *et al.*, 2022). Portanto, destaca-se a necessidade de avaliar a dinâmica de GEE e os processos bioquímicos dos solos relacionados

no setor agrícola do país.

A cultura da cana-de-açúcar é central para a economia e estratégia brasileira, integrando-se ao plano de neo-industrialização através do fomento à bioeconomia. Dessa forma, a Nova Industria Brasil estipula um aumento para 50% de biocombustível na matriz energética (Brasil; Ministério do Desenvolvimento, Indústria, Comércio e Serviços, 2024), além do fator alimentar da açúcar que compõem a cesta básica, revelando a sua importância e justificando a expansão territorial por novas fronteiras no país. No bioma Cerrado, a expansão dessa cultura está associada à irrigação devido ao prolongado período de seca e à aplicação de vinhaça como fertilizante, prática que pode intensificar as emissões de N_2O (Scarpare *et al.*, 2016).

A previsão e o monitoramento das emissões de N_2O em sistemas agrícolas exigem ferramentas capazes de lidar com a não linearidade e a variabilidade espacial e temporal dos dados. Nesse contexto, técnicas de aprendizado de máquina se apresentam como alternativas promissoras, permitindo modelar interações complexas entre variáveis ambientais e de manejo. A aplicação desses modelos possibilita generalizar padrões a partir dos dados existentes, proporcionando uma compreensão mais detalhada das dinâmicas climáticas e operacionais. Dessa forma, tais ferramentas viabilizam novos estudos e aprimoram a precisão das estimativas de GEE na agricultura. Apesar dos avanços na área, observa-se que a literatura ainda carece de análises que abordem a emissão de N_2O sob a perspectiva de como diferentes regimes de irrigação afetam essas emissões e como prever seu comportamento de forma mais precisa. Assim, o presente estudo busca desenvolver e avaliar modelos de aprendizado de máquina para previsão das emissões de N_2O em áreas de cultivo de cana-de-açúcar sob diferentes regimes de irrigação no Cerrado, identificando as variáveis mais influentes no processo.

1.2 Motivação e Justificativa

A compreensão detalhada dos fatores que influenciam as emissões de N_2O é essencial para propor estratégias de mitigação mais eficientes e ambientalmente sustentáveis. O Brasil, como um dos maiores produtores mundiais de cana-de-açúcar, tem papel central na busca por sistemas de alta eficiência produtiva e baixo impacto ambiental.

Embora a irrigação seja fundamental para a produtividade da cana-de-açúcar, seu impacto líquido nas emissões de N2O apresenta resultados muitas vezes contraditórios na literatura recente, atribuindo essa divergência à complexidade dos processos microbianos, onde variáveis como o espaço poroso preenchido por água (WFPS) atuam como 'gatilhos' não-lineares para a emissão (Hamoud; Shaghaleh *et al.*, 2025). Adicionalmente, estudos de modelagem (Leite *et al.*, 2026) sugerem que o balanço final depende intrinsecamente das condições regionais e climáticas, impedindo uma generalização simples sobre o efeito

mitigador ou poluidor da irrigação. Além disso, as abordagens convencionais de análise de dados, baseadas em métodos estatísticos lineares, nem sempre são capazes de capturar a complexidade das interações entre variáveis edafoclimáticas e práticas de manejo agrícola.

O uso de algoritmos de aprendizado de máquina, como *Random Forest* e *Multilayer Perceptron*, pode oferecer previsões mais precisas e identificar variáveis-chave para a mitigação de emissões de GEE, contribuindo para práticas agrícolas mais eficientes. A presente pesquisa busca aprimorar os modelos de analise estatísticos tradicionais, aplicando e comparando modelos de aprendizado de máquina na previsão de fluxos de N_2O em cana-de-açúcar cultivada sob diferentes regimes de irrigação no Cerrado.

1.3 Objetivos

1.3.1 Objetivo Geral

Desenvolver e avaliar modelos de aprendizado de máquina para previsão das emissões de N_2O em áreas de cultivo de cana-de-açúcar sob diferentes regimes de irrigação no Cerrado, identificando as variáveis mais influentes no processo, de modo a consolidar um estimador de tendências e riscos que auxilie na tomada de decisão dos manejos utilizados

1.3.2 Objetivos Específicos

- Avaliar a influência relativa de regimes de irrigação, variáveis de solo e fatores climáticos nas emissões de N_2O.
- Comparar o desempenho dos algoritmos *Random Forest* e *Multilayer Perceptron* na previsão das emissões.
- Identificar variáveis preditoras de maior peso para avaliação de práticas de manejo mais sustentáveis.

1.4 Escopo e Contribuições do Trabalho

Este trabalho está inserido no contexto da agricultura de precisão e da modelagem ambiental, propondo o uso de algoritmos de aprendizado de máquina para prever emissões de N_2O em cana-de-açúcar com diferentes regimes de irrigação no Cerrado. As principais contribuições são:

- Demonstração da aplicabilidade de modelos não lineares na identificação de padrões e riscos de emissão de gases de efeito estufa.
- Identificação de variáveis-chave para mitigação de emissões de N_2O no cultivo de cana.

- Proposição de um *framework* de modelagem escalável e adaptável para outras culturas e regiões.
- Fornecimento de um repositório aberto com dados e código para replicação e extensão dos resultados.

1.5 Organização do Trabalho

O documento está estruturado da seguinte forma:

- **Capítulo 1** – Introdução: apresenta a contextualização do problema, a motivação, os objetivos, o escopo e as contribuições do estudo.
- **Capítulo 2** – Fundamentação Teórica: revisa os conceitos relacionados às emissões de N_2O, à cultura da cana-de-açúcar, ao bioma Cerrado e às técnicas de aprendizado de máquina aplicadas a dados ambientais.
- **Capítulo 3** – Materiais e Métodos: descreve a área de estudo, o delineamento experimental, as variáveis utilizadas, os procedimentos de coleta de dados e os métodos de modelagem.
- **Capítulo 4** – Resultados e Discussão: apresenta e interpreta os resultados obtidos, comparando o desempenho dos modelos e discutindo implicações práticas.
- **Capítulo 5** – Conclusões: sintetiza os principais achados, aponta limitações do estudo e sugere direções para trabalhos futuros.

2 Fundamentação teórica

2.1 Óxido nitroso (N_2O) na agricultura

O N_2O é um gás de efeito estufa de longa permanência na atmosfera, com tempo de vida médio estimado em 120 anos e potencial de aquecimento global 298 vezes maior que o do CO_2. Além de contribuir para o aquecimento global, sendo atualmente o principal responsável pela destruição da camada de ozônio estratosférico.

As emissões de N_2O estão fortemente associadas a atividades agrícolas, principalmente em sistemas que utilizam fertilizantes nitrogenados. Estima-se que aproximadamente 60% das emissões antropogênicas desse gás sejam provenientes da agricultura. O N_2O é produzido principalmente pelos processos microbianos de nitrificação e desnitrificação, que são influenciados por fatores do solo como temperatura, pH, disponibilidade de nitrogênio, teor de umidade, além do manejo de fertilizantes e irrigação.

2.2 Cana-de-açúcar no Brasil e no Cerrado

A cana-de-açúcar (*Saccharum officinarum*) é uma das principais culturas agrícolas do Brasil, ocupando mais de 8 milhões de hectares. No bioma Cerrado, a cana-de-açúcar tem apresentado grande expansão devido à disponibilidade de áreas e à crescente demanda por biocombustíveis, como o etanol.

O cultivo de cana-de-açúcar nessa região enfrenta, contudo, desafios relacionados ao regime climático. O Cerrado apresenta uma estação seca bem definida, com déficit hídrico de abril a setembro. Para manter a produtividade da cana, a irrigação é frequentemente utilizada, o que modifica a dinâmica de água no solo e, consequentemente, pode impactar favorecendo emissão de N_2O.

Além disso, práticas como a aplicação de vinhaça — resíduo líquido resultante da produção de etanol — aumentam a disponibilidade de carbono e nitrogênio no solo, potencializando os processos de nitrificação e desnitrificação e elevando as emissões de N_2O. Assim, compreender a relação entre irrigação, fertilização e emissões de N_2O é essencial para propor práticas agrícolas mais sustentáveis.

2.3 Regimes de irrigação e emissões

As diferentes estratégias de irrigação podem alterar significativamente os fluxos de N_2O em solos cultivados. A irrigação de salvamento, aplicada em situações esporádicas

de déficit hídrico, como veranicos, tende a provocar aumentos pontuais de umidade no solo, frequentemente associados a picos de emissão de N_2O devido à rápida ativação de microrganismos, ou seja, as bactérias nitrificantes que atuam nos processos de nitrificação e denitrificação.

Por outro lado, sistemas de reposição contínua, ajustados a 17%, 46% e 75% da evapotranspiração da cultura (ETc), modularam a disponibilidade de água no solo de maneira distinta. Níveis mais elevados de reposição (75% da ETc) favorecem condições de maior saturação hídrica e, portanto, intensificam a produção de N_2O. Já regimes intermediários (46% da ETc) podem representar um equilíbrio entre a manutenção da produtividade agrícola e a mitigação das emissões de N_2O.

A aplicação de vinhaça como insumo agrícola adiciona carbono e nitrogênio ao solo, funcionando como substrato adicional para os microrganismos envolvidos na produção de N_2O e na presença de umidade. Quando combinada a regimes de irrigação elevados, essa prática pode aumentar substancialmente o risco de intensificação das emissões de N_2O, ressaltando a necessidade de práticas de manejo adequadas.

2.4 Modelagem de emissões de N_2O

2.4.1 Modelos tradicionais

Tradicionalmente, modelos estatísticos lineares têm sido utilizados para estimar as emissões de N_2O em diferentes sistemas agrícolas. Embora úteis, esses modelos muitas vezes não capturam a complexidade e a não linearidade das interações entre variáveis edafoclimáticas e práticas de manejo.

A literatura mostra que as emissões de N_2O apresenta alta variabilidade temporal e espacial, com fluxos intensos em períodos curtos após a adubação ou eventos de chuva/irrigação. Essa característica dificulta a modelagem com métodos convencionais e sugere a necessidade de técnicas mais flexíveis e adaptativas.

2.4.2 Aprendizado de máquina

O aprendizado de máquina é uma subárea da inteligência artificial que utiliza algoritmos capazes de aprender padrões a partir de dados. Diferente dos modelos estatísticos tradicionais, que partem de hipóteses paramétricas, os modelos de aprendizado de máquina podem lidar com não linearidades, variáveis correlacionadas e grandes conjuntos de dados com relativa eficiência.

Entre os algoritmos mais aplicados em estudos ambientais estão:

- **Random Forest (RF)**: método baseado em múltiplas árvores de decisão que combina suas predições para aumentar a robustez e reduzir o sobreajuste.
- **Redes Neurais Artificiais (RNA)**: estruturas computacionais inspiradas no funcionamento do cérebro humano, capazes de representar relações complexas entre variáveis.
- **Support Vector Machines (SVM)**: algoritmos que buscam encontrar hiperplanos de separação entre classes ou funções de regressão, aplicados em problemas de previsão.

Diversos estudos recentes têm utilizado aprendizado de máquina para prever fluxos de gases de efeito estufa em diferentes sistemas agrícolas, com resultados superiores aos obtidos por métodos lineares. Esses modelos também possibilitam avaliar a importância relativa das variáveis preditoras, oferecendo subsídios para a tomada de decisão em práticas de manejo agrícola sustentável.

2.5 Trabalhos relacionados

Diversos estudos têm investigado a variabilidade das emissões de N_2O em sistemas agrícolas, destacando-se especialmente aqueles que analisam a interação entre disponibilidade de nitrogênio mineral, umidade do solo, temperatura e manejo a partir de diferentes abordagens teóricas e metodológicas. Os trabalhos tradicionais na área concentram-se em modelos baseados em processos (process-based models) como DNDC (Li *et al.*, 2005), DAYCENT (Parton *et al.*, 2001) e SWAT (Arnold *et al.*, 1998), buscando explicar a nitrificação e desnitrificação com base na interpretação mecanística desses processos.

Embora tais modelos apresentem limitações em ambientes tropicais devido à sensibilidade de parametrização, estudos recentes demonstram avanços em sua aplicação local. Silva *et al.* (2024), por exemplo, observaram que o desempenho do modelo STICS foi adequado tanto para o sistema de manejo convencional quanto para o plantio direto no Cerrado. Ao projetar cenários climáticos para o período de 2021 a 2070, os autores identificaram fortes indícios de redução na produtividade de grãos e na biomassa aérea total. Contraditoriamente à queda de produtividade, observou-se uma tendência de aumento nas emissões de N_2O em ambos os sistemas de manejo, fenômeno atribuído ao aumento da temperatura e à redução do ciclo da cultura induzidos pelas mudanças climáticas.

Paralelamente à modelagem de processos, autores têm adotado abordagens orientadas a dados, investigando o uso de aprendizado de máquina por meio de algoritmos como Random Forest e Gradient Boosting (Saha; Basso; Robertson, 2021). Essas técnicas ampliam o entendimento sobre as interações complexas do solo. Enquanto estudos em cana-de-açúcar irrigada evidenciam o papel central do regime hídrico — com maiores fluxos de N_2O associados a elevados WFPS e disponibilidade de NH_4^+ e NO_3^- (Carvalho *et al.*, 2021) —, resultados

obtidos via Random Forest convergem com o presente estudo ao ratificar o NH_4^+ e o NO_3^- como variáveis preditoras determinantes.

Nesse contexto, o presente trabalho contribui ao preencher essas lacunas a partir da análise de dados experimentais com métodos de aprendizado de máquina, destacando-se por modelos de ampla aceitação científica com base em seus desempenhos, oferecendo uma abordagem que dialoga com a literatura, mas avança ao incorporar análise de importância por permutação, curvas ICE e avaliação comparativa entre conjunto desbalanceado e conjunto balanceado, algo ainda pouco explorado na predição de N_2O.

3 Materiais e métodos

3.1 Local do experimento

Os dados utilizados são resultado de um estudo realizado em 2010 na estação experimental da Empresa Brasileira de Pesquisa Agropecuária (Embrapa – Cerrados)(Fig.1), localizada em Planaltina, DF, Brasil (15°36′17.76″S 47°42′35.51″W) (Carvalho *et al.*, 2021).

Figura 1 – Embrapa Cerrados,Distrito Federal,Brasil (Carvalho *et al.*, 2021).

O experimento foi conduzido em área de cultivo de cana-de-açúcar localizada no bioma Cerrado, região caracterizada por clima Aw-tropical chuvoso (Köppen). O microclima da área de estudo possui temperatura média anual entre 22 °C e 23 °C, e precipitação média anual de 1383 mm, concentrada entre outubro e março. O solo é classificado como Latossolo, de textura argilosa. Foi realizada análise química do solo na camada de 0 a 20 cm, conforme apresentado na Tabela 1.

A adubações na implantação do experimento consistiu na aplicação de 4 t ha^{-1} de calcário e 500 kg ha^{-1} de gesso, além de 50 kg ha^{-1} de FTE BR-10 (2,5% B, 1,0 Cu, 4% Mn, 0,1% Mo, 4% Fe, 7% Zn e 0,1% Co).

A variedade de cana-de-açúcar utilizada foi a RB855536, plantada em 18 de junho de 2010 com 600 kg ha^{-1} de NPK (4-30-16), e FTE–BR 12 como fonte de micronutrientes (Zn, B, Cu, Fe, Mn e Mo, com 7,0; 2,5; 1,0; 4,0; 4,0 e 0,1%, respectivamente). Utilizou-se espaçamento de 1,5 metros entre plantas.

As condições climáticas foram monitoradas por estação meteorológica automática instalada próxima à área experimental, registrando dados diários de precipitação, temperatura do ar, umidade relativa e radiação solar.

Tabela 1 – Análise química da camada de solo de 0-20 cm utilizada neste estudo (Carvalho *et al.*, 2021).

Componente	Valor
pH (H_2O)	5,08
Alumínio (Al^{3+})	0,39 cmolc dm^{-3}
Fósforo (P)	0,22 mg dm^{-3}
Potássio (K^+)	8,0 mg dm^{-3}
Cálcio (Ca^{2+})	0,56 cmolc dm^{-3}
Magnésio (Mg^{2+})	0,26 cmolc dm^{-3}
Acidez potencial (H+Al)	3,7 cmolc dm^{-3}
Matéria orgânica	0,87 $g.kg^{-1}$

3.2 Delineamento experimental

O experimento foi conduzido utilizando quatro diferentes tipos de regimes de irrigação, com repetição de cada tratamento em três blocos, totalizando 12 parcelas experimentais. Cada bloco possuía duas linhas de plantio com 4 metros de comprimento. O tratamento com menor volume de irrigação foi o regime de irrigação denominada de salvamento (R - *Rescue*), no qual foi aplicada apenas uma irrigação de 40 mm para compensar o défit hídrico do solo. Nos demais regimes, a reposição hídrica foi estabelecida de acordo com um percentual da demanda da cultura, sendo 17% (T17%), 46% (T46%) e 75% (T75%) da evapotranspiração da cultura (ETc) conforme a Figura 3.2, calculada pelo método de Penman–Monteith e ajustada pelo coeficiente de evapotranspiração da cana-de-açúcar (P.R. da Silva *et al.*, 2013).

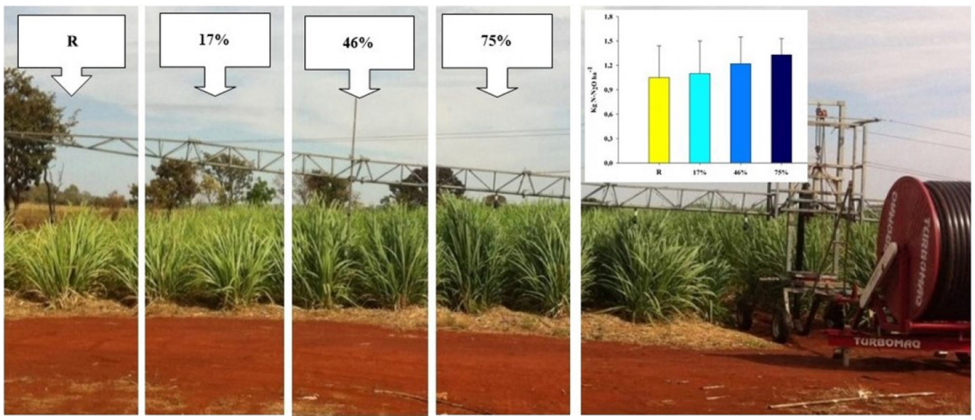

Figura 2 – Área experimental, Embrapa Cerrados,Distrito Federal,Brasil (Carvalho *et al.*, 2021).

O sistema de irrigação utilizado foi por aspersores de velocidade ajustável para diferenciar cada regime, conectados a uma barra irrigadora de 40 metros com 10 aspersores de cada lado.

Foram utilizadas duas câmaras estáticas fechadas em cada parcela, posicionadas em paralelo à linha de plantio: uma entre as duas linhas do mesmo bloco e a outra mais próxima da linha. No total, foram utilizadas 24 câmaras estáticas no campo experimental, e

as coletas foram realizadas entre 22 de junho de 2015 e 29 de maio de 2016. Adicionalmente, foram coletadas três amostras de fluxo de N_2O no mesmo tipo de solo com vegetação nativa de Cerrado (CE). Todos os dados, como mencionado anteriormente, são originalmente de (Carvalho *et al.*, 2021).

Tabela 2 – Volume de água (mm) aplicado durante os eventos de irrigação na área experimental, Embrapa Cerrados, Distrito Federal, Brasil (Carvalho *et al.*, 2021).

Data de irrigação	**mm de água aplicada**			
	R	**T17%**	**T46%**	**T75%**
13/jun/2015	40.0	40.0	40.0	40.0
23/jun/2015	0	6.1	24.6	32.9
02/jul/2015	0	3.1	11.4	22.3
13/jul/2015	0	4.8	15.5	27.9
23/jul/2015	0	6.3	25.7	35.5
03/ago/2015	0	7.3	26.1	32.2
12/ago/2015	0	4.5	15.1	19.8
24/ago/2015	0	7.5	20.9	21.8
03/set/2015	0	6.3	16.2	35.5
14/set/2015	0	8.0	28.6	33.3
24/set/2015	0	3.8	11.9	20.1
05/out/2015	0	6.8	18.7	34.9
15/out/2015	0	2.4	9.2	17.5
11/nov/2015	0	2.5	6.8	18.3
Total	40.0	109.0	270.7	392.0

R – "Irrigação de salvamento" de 40 mm, aplicada logo após o corte da cana-de-açúcar.

O delineamento adotado foi em blocos casualizados, com diferentes regimes de irrigação aplicados às parcelas experimentais de cana-de-açúcar. Os tratamentos incluíram:

- **Irrigação de salvamento salvamento (R - *Rescue*)**: aplicação de água apenas em condições de déficit hídrico severo.
- **Reposição de 17% da ETc**: reposição parcial da evapotranspiração da cultura.
- **Reposição de 46% da ETc**: regime intermediário de irrigação.
- **Reposição de 75% da ETc**: regime de maior reposição hídrica.

Além dos quatro regimes, foi avaliada um área de cerrado nativo adjacente ao campo experimental, sendo utilizado como área de referencia durante o monitoramento das emissões de N_2O. Como recomendado pelo Painel Intergovernamental sobre Mudança do Clima - IPCC (Intergovernmental Panel on Climate Change (IPCC), 2006).

3.3 Coleta de dados e variáveis medidas

As emissões de N_2O foram medidas utilizando o método de câmaras estáticas acopladas a cromatógrafo a gás. As coletas foram realizadas em intervalos regulares, com maior frequência após eventos de irrigação e aplicação de fertilizantes.

Para cada regime de irrigação, foram instaladas câmaras em pontos representativos da parcela, considerando bordadura e posição central. As concentrações de N_2O foram quantificadas e os fluxos calculados a partir da taxa de variação da concentração em função do tempo (Carvalho *et al.*, 2021).

Tabela 3 – Variáveis experimentais, acrônimos e unidades.

Acrônimo	**Significado (unidade)**
NO_3^-	Nitrato (mg kg^{-1})
NH_4^+	Amônio (mg kg^{-1})
WFPS	Espaço poroso preenchido por água (%)
ST	Temperatura do solo (ºC)
Tirriga	Nível de irrigação (% ETc)
Precip	Precipitação (mm)
Tmed	Temperatura média (ºC)
Avg. RH	Umidade relativa média (%)
Wind	Velocidade do vento (m s^{-1})
Insol	Insolação (h)
Rad	Radiação (MJ m^{-2} dia^{-1})
Evap	Evapotranspiração potencial (mm)
N_2O	Fluxo de N_2O (μg m^{-2} h^{-1})

Essas variáveis foram selecionadas por sua relevância na literatura e disponibilidade no experimento, permitindo capturar a interação entre condições edafoclimáticas e práticas de manejo.

3.4 Pré-processamento dos dados

Os dados coletados foram organizados em planilhas e submetidos a pré-processamento. Foram realizadas as seguintes etapas:

- Tratamento de valores faltantes por interpolação ou exclusão de registros incompletos.
- Normalização das variáveis contínuas para reduzir diferenças de escala.
- Criação de um conjunto de dados **desbalanceado**, refletindo as frequências reais de emissão observadas.

- Criação de um conjunto de dados **balanceado**, obtido com o uso da função `pd.qcut`, gerando classes de igual tamanho para treinamento dos algoritmos.

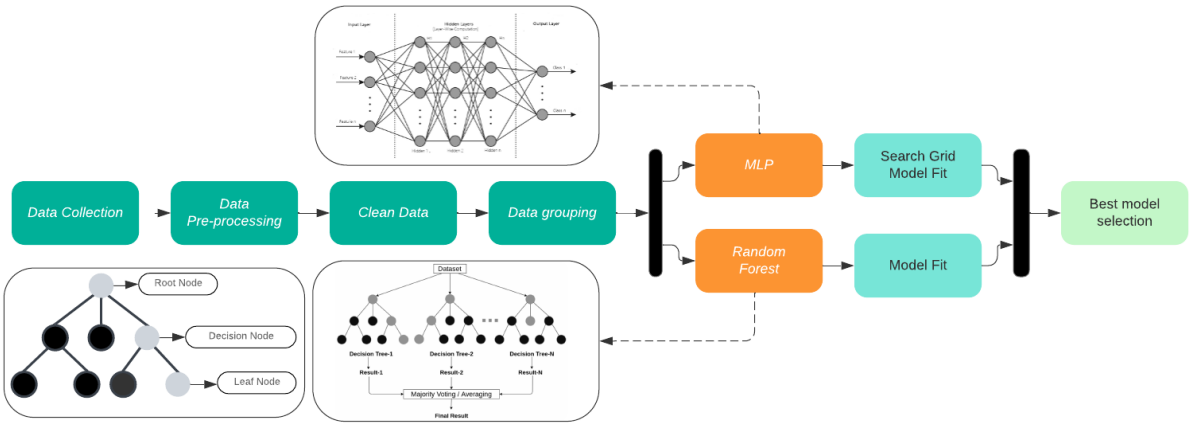

Figura 3 – Fluxograma de análise de dados (*flowchart*).

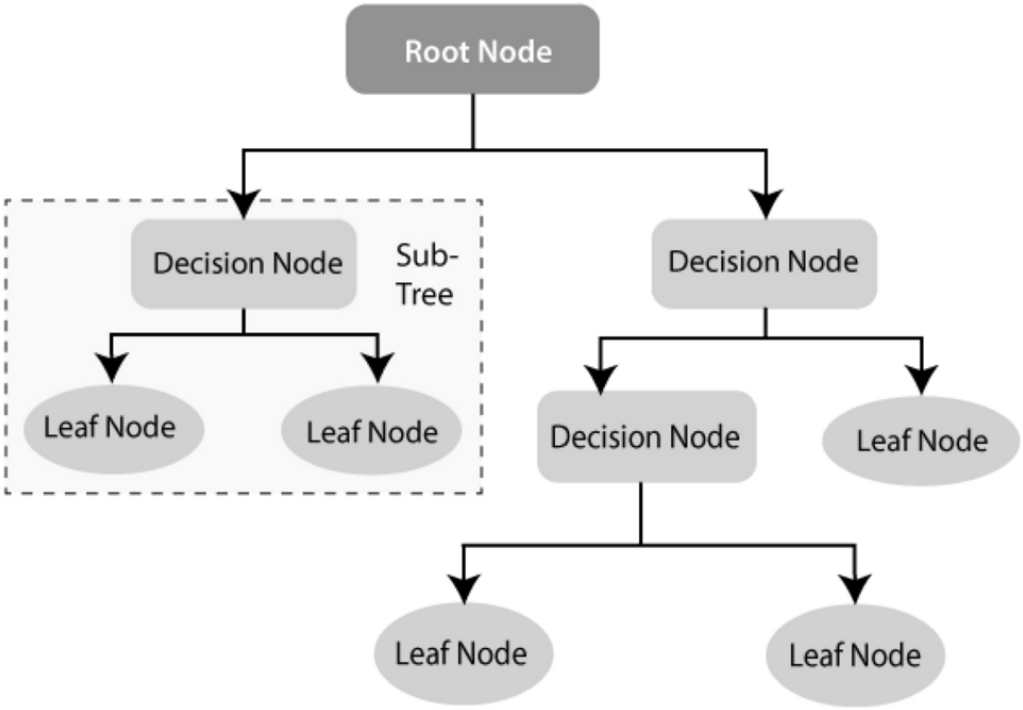

Figura 4 – Estrutura típica de uma árvore de decisão (DT).

3.4.1 Delimitação das Contribuições

As análises agronômicas e a coleta de dados de campo utilizadas nesta dissertação baseiam-se em protocolos consolidados em publicações prévias de 2021 (Carvalho *et al.*, 2021). A contribuição original deste trabalho reside no desenvolvimento do arcabouço computacional, na arquitetura dos modelos de ML e na implementação das técnicas de interpretabilidade (ICE e Permutação) aplicadas a este dataset.

3.4.2 Considerações sobre Viabilidade Técnica

Embora o modelo dependa de variáveis como NO_3^- e NH_4^+, cujas medições atuais são laboratoriais, a modelagem proposta fundamenta a lógica para futuros sistemas mecatrônicos integrados a sensores de íons em tempo real, permitindo a automação do manejo hídrico baseada em riscos.

3.5 Modelagem por aprendizado de máquina

O aprendizado de máquina constitui uma abordagem computacional orientada por dados que possibilita a realização de previsões com elevado grau de precisão e confiabilidade. Trata-se de um conjunto de técnicas particularmente eficaz para a exploração e análise de grandes volumes de dados complexos e de alta dimensionalidade. Esses métodos podem ser classificados de acordo com seus paradigmas de aprendizagem. Na aprendizagem não supervisionada, os algoritmos operam sobre dados não rotulados, buscando identificar padrões, estruturas ou agrupamentos intrínsecos ao conjunto de dados. Em contrapartida, na aprendizagem supervisionada, os dados de entrada são previamente rotulados, permitindo o mapeamento das variáveis explicativas em relação a uma variável-alvo conhecida.

Entre os métodos de aprendizagem supervisionada amplamente utilizados para avaliação de desempenho, destacam-se o Random Forest (RF) e as redes neurais do tipo Multilayer Perceptron (MLP), reconhecidos por sua elevada eficácia na modelagem de relações não lineares. Em razão dessas características, ambos os algoritmos foram selecionados para este estudo. A Figura 3 apresenta o fluxograma geral da metodologia adotada.

3.5.1 Random Forest (RF)

O algoritmo Random Forest é composto por um número definido de árvores de decisão (Decision Trees – DT), configurando-se como uma ferramenta versátil, capaz de produzir múltiplas soluções para um mesmo problema. Cada árvore de decisão é construída a partir de um nó raiz e se ramifica em subárvores formadas por nós de decisão e nós folha (Breiman, 2001). A Figura 4 apresenta essa estrutura hierárquica que é resultante de um processo recursivo de particionamento dos dados, no qual o conjunto original é sucessivamente dividido em subconjuntos mais homogêneos.

Em cada nó da árvore, uma função de separação determina a melhor forma de dividir os dados, geralmente baseada em medidas como entropia ou impureza de Gini. Esse procedimento é repetido ao longo de toda a árvore, definindo intervalos de segmentação progressivamente mais específicos até a obtenção dos nós folha, nos quais os subconjuntos de dados apresentam o maior grau possível de uniformidade. O número de amostras em cada nó folha pode variar, desde um único ponto de dado até múltiplas observações, dependendo

da profundidade e da complexidade da árvore. O crescimento das árvores é controlado por critérios de parada previamente estabelecidos, que limitam sua expansão estrutural.

A partir desse processo, o modelo calcula métricas de importância das variáveis, permitindo avaliar a contribuição de cada variável de entrada nas previsões realizadas, além de identificar interações não lineares entre os atributos. Os dados são apresentados ao modelo em subconjuntos, com repetição em diferentes execuções paralelas de treinamento das árvores de decisão. Cada árvore produz uma previsão individual, e os resultados são posteriormente combinados por meio de um processo conhecido como majority voting, no qual a resposta mais frequente é adotada como previsão final. Esse mecanismo contribui para a redução do ruído presente nos dados e aumenta a robustez do modelo. Dessa forma, o Random Forest é amplamente empregado em problemas de classificação e regressão (Belgiu; Drăguţ, 2016; Islam *et al.*, 2021; Fan *et al.*, 2022).

3.5.2 Multilayer Perceptron (MLP)

O MLP é um tipo de rede neural artificial composta por múltiplas camadas de nós interconectados, denominados neurônios, sendo capaz de aprender representações complexas dos dados (Fischer, 2015). A arquitetura mínima de uma MLP inclui três camadas: uma camada de entrada, que recebe os valores das variáveis explicativas; uma ou mais camadas ocultas intermediárias; e uma camada de saída, responsável por fornecer as previsões associadas à variável-alvo. A interação entre essas camadas é inlustrada na Figura 5.

O funcionamento de cada neurônio é determinado por um conjunto de valores de entrada, que são processados por meio de funções de ativação, resultando em um valor de saída. Essas funções de ativação são não lineares, o que possibilita à rede neural modelar relações complexas e não lineares presentes nos dados. Os neurônios de uma camada estão conectados aos da camada subsequente, sendo que cada conexão possui um peso associado, ajustado durante o processo de aprendizagem. O treinamento da rede ocorre a partir da apresentação de um conjunto de dados à camada de entrada, e os erros resultantes são utilizados por um algoritmo de otimização em conjunto com a técnica de Backpropagation, que propaga o erro da camada de saída para as camadas anteriores, promovendo o ajuste iterativo dos pesos das conexões (Alpaydın, 2004).

Para a otimização da arquitetura da rede MLP, foi empregado o algoritmo de busca exaustiva GridSearch. Esse método avalia sistematicamente todas as combinações possíveis de um conjunto predefinido de hiperparâmetros, identificando a configuração que apresenta o melhor desempenho de acordo com uma métrica de avaliação específica. Os parâmetros considerados nesta simulação incluíram arquiteturas de camadas ocultas [(50, 50, 5) e (100, 100, 5)], funções de ativação (relu e tanh), algoritmos de otimização (adam e lbfgs), estratégias de taxa de aprendizado (constante ou adaptativa), com valores variando de 0,1 a 0,001 em potências de 10, além do parâmetro de regularização alpha, também variando de 0,1 a 0,001

em múltiplos de 10. O número máximo de iterações de treinamento foi fixado em 2000, de modo a garantir uma avaliação eficiente das diferentes configurações testadas.

Os modelos foram treinados e testados utilizando validação cruzada, de modo a evitar sobreajuste e garantir maior generalização. As métricas utilizadas para avaliar o desempenho dos algoritmos foram:

- Coeficiente de determinação (R^2).
- Erro quadrático médio (RMSE).
- Erro absoluto médio (MAE).

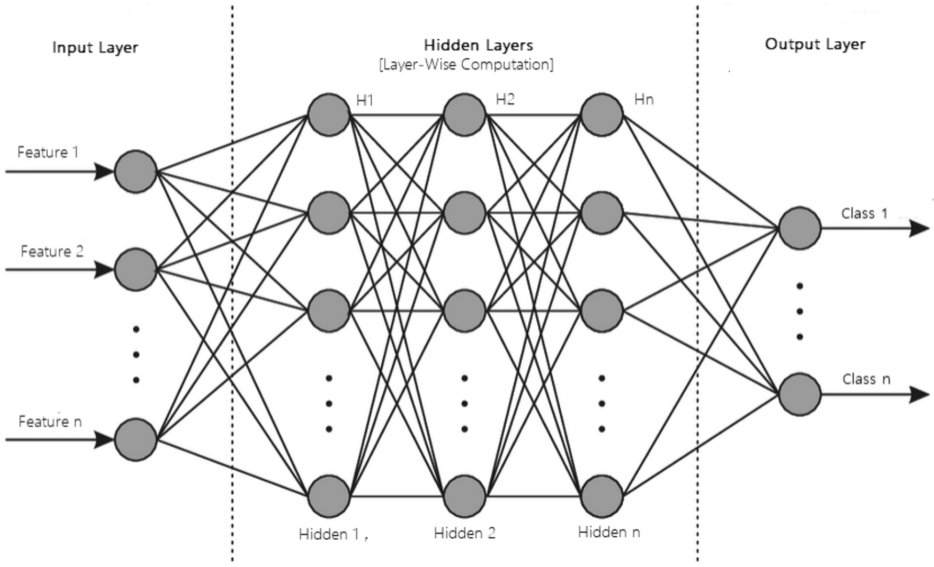

Figura 5 – Arquitetura de uma rede neural perceptron multicamadas (MLP).

3.6 Métricas de avaliação

A principal métrica adotada para a validação dos algoritmos foi o erro médio absoluto (*Mean Absolute Error* – MAE), calculado a partir da média das diferenças absolutas entre os valores previstos e os valores observados. Essa métrica fornece uma estimativa direta do erro médio das previsões, expressa na mesma unidade da variável analisada. Como métrica complementar, utilizou-se o erro médio quadrático (*Mean Squared Error* – MSE), que considera o quadrado das diferenças entre os valores previstos e observados, tornando-se mais sensível a erros de maior magnitude e penalizando previsões imprecisas de forma mais acentuada (Alpaydın, 2004). A partir do MSE, obtém-se o erro quadrático médio da raiz (*Root Mean Squared Error* – RMSE), definido como a raiz quadrada do erro médio quadrático, permitindo que o erro seja interpretado novamente na unidade original dos dados.

Adicionalmente, foi empregado o coeficiente de determinação (R^2), que indica a proporção da variância total dos dados observados explicada pelo modelo, sendo amplamente utilizado para avaliar a qualidade do ajuste em problemas de regressão.

As Equações (3.1) a (3.4) apresentam as expressões matemáticas utilizadas para o cálculo dessas métricas:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{3.1}$$

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{3.2}$$

$$\text{RMSE} = \sqrt{\text{MSE}} \tag{3.3}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{3.4}$$

onde n representa o número total de observações, y_i corresponde ao valor real do i-ésimo dado, $\hat{y}_i$ indica o valor previsto pelo modelo para esse ponto, e $\bar{y}$ representa a média dos valores observados.

3.7 Ambiente Computacional

A modelagem foi realizada em ambiente Python, utilizando bibliotecas de ciência de dados como `scikit-learn`, `pandas`, `numpy` e `matplotlib`. O código desenvolvido e utilizado neste estudo está disponível publicamente para consulta, verificação e reutilização. O repositório pode ser acessado na plataforma GitHub, no endereço: https://github.com/rafaeltb/Prediction-of-sugarcane-N2O-emissions . A disponibilização do código visa assegurar a reprodutibilidade dos resultados apresentados, bem como possibilitar que outros pesquisadores utilizem, adaptem e ampliem as metodologias propostas neste trabalho.

4 Resultados e discussão

4.1 Análises exploratórias

As emissões de N_2O medidas nos tratamentos T75%, T46%, T17%, R e na vegetação nativa de Cerrado apresentaram amplitudes que variaram, respectivamente, de −35,7 a 117, de −84,6 a 47,8, de −26,7 a 92,7, de −13,2 a 35,1 e de −406 a 61,9 $\mu g\, m^{-2}\, h^{-1}$ de N_2O. Os maiores valores de emissão foram observados no tratamento T75%, enquanto os menores ocorreram nos tratamentos T46% e R.

As relações entre os regimes de irrigação, as condições de referência e as variáveis N_2O, NO_3^-, NH_4^+ e temperatura do solo (TS) são apresentadas na Fig. 6, juntamente com seus valores médios. As interações entre as variáveis foram inicialmente avaliadas por meio da construção de uma matriz de correlação (Fig. 7). Observou-se que as correlações estatisticamente significativas estiveram predominantemente associadas aos parâmetros climáticos. Em contrapartida, as interações envolvendo as emissões de N_2O apresentaram correlações fracas, indicando a ausência de um único preditor dominante entre as variáveis mensuradas.

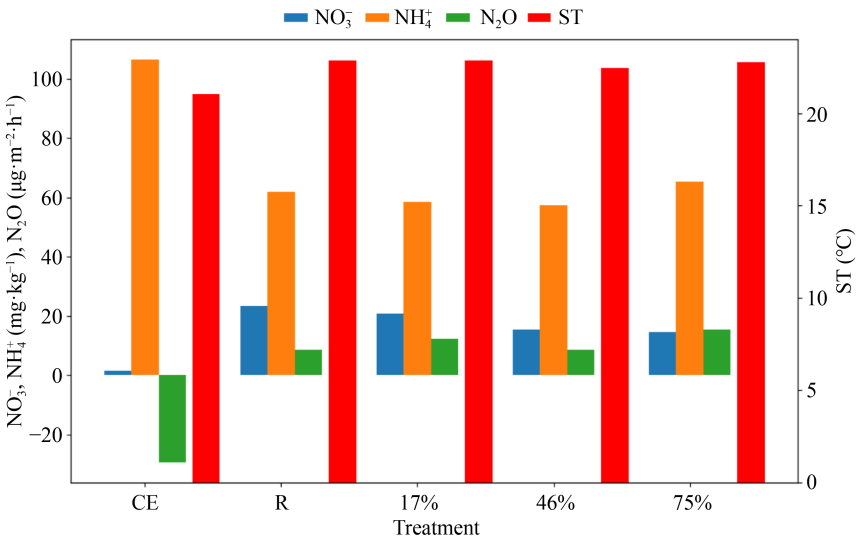

Figura 6 – Gráfico dos valores médios de N_2O, NO_3^-, NH_4^+ e temperatura do solo (TS) obtidos nos diferentes tratamentos: Cerrado (CE), salvamento(R) e irrigação nos níveis T17%, T46% e T75%.

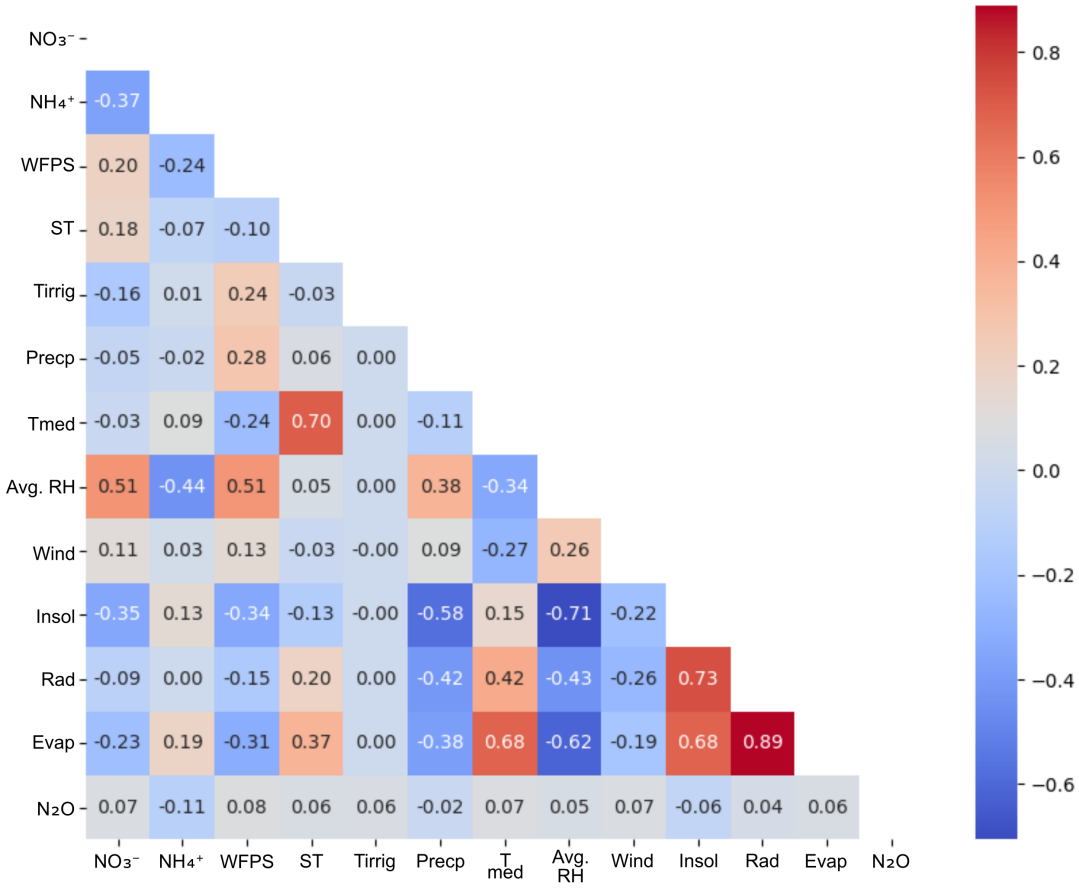

Figura 7 – Heatmap de correlação entre variáveis do estudo.

4.2 Balanceamento de dados

Dentro desse cenário, os dados em estudo apresentam elevada heterogeneidade na distribuição dos valores, o que motivou a organização do conjunto amostral em dez grupos distintos (G). Inicialmente, os modelos foram aplicados utilizando os dados em sua forma contínua e original. Em seguida, os dados foram distribuídos em grupos desbalanceados, definidos a partir de intervalos de valores com amplitudes iguais. Por fim, adotou-se uma estratégia de balanceamento, na qual cada grupo passou a conter a mesma quantidade de observações, resultando em intervalos de valores não uniformes. Essa configuração evidencia a necessidade de avaliar o impacto da distribuição dos dados no desempenho dos modelos preditivos, conforme ilustrado na Figura 8.

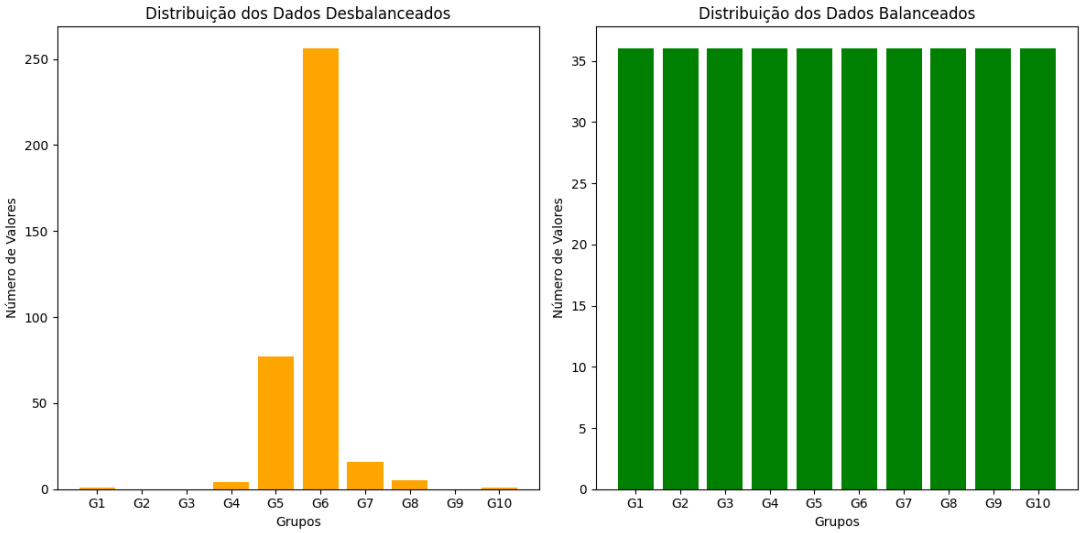

Figura 8 – Conjuntos desbalanceado e balanceado.

Uma vez que os dados analisados apresentam desbalanceamento entre os grupos quando organizados por intervalos uniformes, também foi investigado o efeito da aplicação de uma estratégia de balanceamento baseada na redistribuição das observações. Diferentemente de abordagens clássicas de balanceamento por amostragem, neste estudo o balanceamento foi realizado por meio da redefinição dos limites dos intervalos, de forma a garantir o mesmo número de observações em cada grupo.

Assim, duas abordagens principais de organização dos dados agrupados foram consideradas: (i) o agrupamento desbalanceado, no qual os intervalos de valores são uniformes, mas a quantidade de observações por grupo varia significativamente; e (ii) o agrupamento balanceado, no qual a quantidade de observações por grupo é mantida constante, ao custo de intervalos de valores com amplitudes distintas.

Ressalta-se que, para ambos os conjuntos agrupados os valores máximos e mínimos de cada grupo foram previamente definidos e permanecem fixos ao longo das análises, conforme apresentado na Tabela 4. Essa estratégia assegura a comparabilidade entre os experimentos e permite avaliar exclusivamente o impacto da forma de distribuição dos dados sobre o desempenho dos modelos.

4.3 Desempenho dos modelos

As emissões de N_2O foram estimadas por meio de dois modelos distintos de *machine learning*: *Random Forest* (RF) e *Multilayer Perceptron* (MLP). O modelo RF foi desenvolvido a partir dos valores médios obtidos nos três blocos experimentais do estudo, considerando variáveis relacionadas ao solo, ao clima e aos gases. Em paralelo, o modelo MLP foi treinado utilizando o mesmo conjunto de dados, com otimização de hiperparâmetros realizada por meio do algoritmo *GridSearch*, visando identificar a arquitetura de rede neural mais eficiente.

Tabela 4 – Intervalos dos grupos para os conjuntos desbalanceado e balanceado

Grupo	Desbalanceado	Balanceado
G1	-255.4 a -203.8	-254.9 a -3.36
G2	-203.8 a -152.7	-3.36 a 0.12
G3	-152.7 a -101.6	0.12 a 1.83
G4	-101.6 a -50.5	1.83 a 3.61
G5	-50.5 a 0.59	3.61 a 5.90
G6	0.59 a 51.69	5.90 a 8.51
G7	51.69 a 102.79	8.51 a 13.70
G8	102.79 a 153.88	13.70 a 20.30
G9	153.88 a 204.98	20.30 a 35.01
G10	204.98 a 256.08	35.01 a 256.08

Do total de 390 valores medidos de emissões de N_2O, observou-se a presença de agrupamento dos dados, caracterizado por uma maior concentração de observações em determinados intervalos de valores. Para lidar com esse desbalanceamento, foram construídos três cenários experimentais: um conjunto de dados não agrupado, um conjunto desbalanceado contendo todas as medições originais e um conjunto balanceado, composto por 10 subconjuntos com 36 medições cada, distribuídas uniformemente ao longo de intervalos padronizados de valores. Em todos os cenários, os modelos selecionaram aleatoriamente 80% dos dados para treinamento, utilizando os 20% restantes para validação.

A aplicação dos algoritmos de aprendizado de máquina revelou diferenças significativas no desempenho preditivo entre os métodos avaliados. Os resultados detalhados do desempenho do modelo RF nos diferentes cenários são apresentados na Tabela 5. Observa-se que o conjunto de dados balanceado apresentou o melhor desempenho preditivo, com $R^2 = 0{,}8736$, enquanto o conjunto desbalanceado obteve o menor valor de coeficiente de determinação. É imperativo destacar que, no cenário de dados não agrupados, o erro (MAE) do modelo RF é significativamente superior (aproximadamente 12 vezes maior) ao cenário balanceado. Este comportamento reforça que a acurácia do modelo é altamente dependente do pré-processamento estatístico e do balanceamento das classes de emissão.

A comparação entre os valores observados e preditos pelo modelo RF é apresentada na Fig. 9. A análise dos pontos revela que a tendência dos dados não atinge o ângulo ideal de 45° representado pela linha pontilhada. Esta inclinação evidencia a limitação do Random Forest em capturar valores extremos (picos de emissão) em conjuntos de dados limitados, tendendo a subestimar magnitudes elevadas.

Tabela 5 – Métricas do modelo RF em diferentes cenários.

Métrica	Não agrupado	Desbalanceado	Balanceado
MAE	4,7914	0,1093	0,3894
MSE	73,0465	0,0462	0,2546
RMSE	8,5467	0,2150	0,5045
R^2	0,8216	0,8094	0,8736

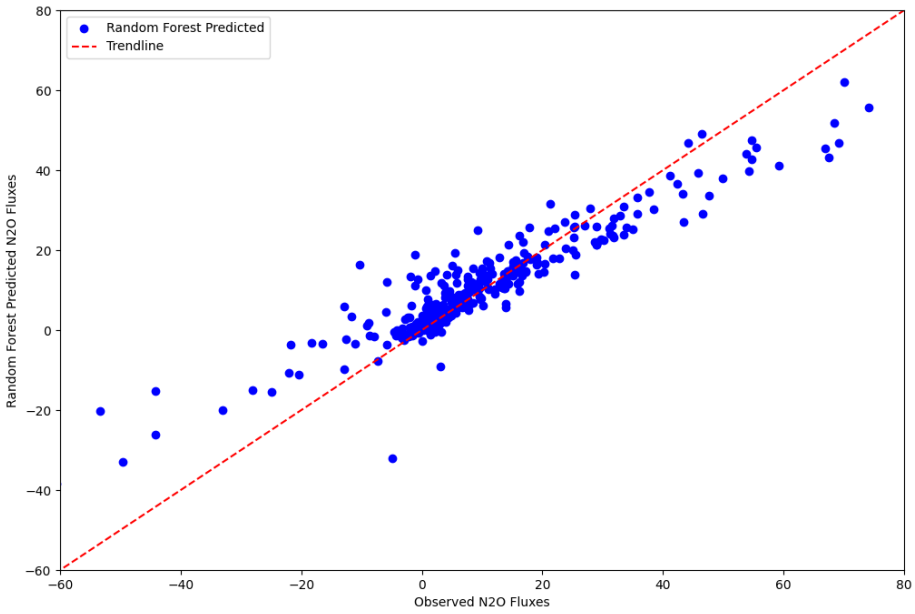

Figura 9 – Comparação entre valores observados e preditos de emissões de N_2O utilizando o modelo RF.

Em contraste, o modelo *Multilayer Perceptron*, embora apresente capacidade teórica para modelar relações não lineares complexas, obteve desempenho inferior ao RF em todos os cenários avaliados. O melhor resultado do MLP foi observado no conjunto de dados balanceado, com coeficiente de determinação de 53,60%. A menor acurácia do modelo pode estar associada à necessidade de um maior volume de dados para o treinamento eficiente de redes neurais artificiais, bem como à sua sensibilidade aos parâmetros de regularização e às taxas de aprendizado.

As métricas de desempenho do modelo MLP são apresentadas na Tabela 6, enquanto a comparação entre os valores observados e preditos é ilustrada na Fig. 10. A partir do processo de otimização de hiperparâmetros utilizando o algoritmo *GridSearch* em conjunto com as redes neurais MLP, foi possível identificar as configurações mais influentes para cada conjunto de dados. Para o conjunto balanceado, a configuração ótima incluiu a função de ativação *tanh*, camadas ocultas compostas por 100, 100 e 5 neurônios, respectivamente, taxa de aprendizado constante com valor inicial de 0,001, número máximo de 2000 iterações e o

solucionador *adam*.

Por outro lado, o conjunto de dados desbalanceado apresentou melhor desempenho com a mesma função de ativação *tanh*, porém com hiperparâmetros distintos, incluindo camadas ocultas de 50, 50 e 5 neurônios, respectivamente, taxa de aprendizado constante, valor inicial de taxa de aprendizado de 0,1, número máximo de 2000 iterações e o solucionador *adam*. De forma semelhante, o conjunto de dados não agrupado obteve resultados ótimos utilizando a função de ativação *tanh*, camadas ocultas de 50, 50 e 5 neurônios, taxa de aprendizado constante, valor inicial de 0,1, limite de 2000 iterações e o solucionador *adam*.

Essa configuração otimizada resultou em uma melhoria na acurácia preditiva dos modelos, evidenciando a importância de ajustar os hiperparâmetros de acordo com as características específicas de cada conjunto de dados.

Tabela 6 – Métricas do modelo MLP em diferentes cenários.

Métrica	**Não agrupado**	**Desbalanceado**	**Balanceado**
MAE	1,2574	1,2765	0,5682
MSE	2,2065	2,2320	0,8999
RMSE	1,4854	1,4940	0,9486
R^2	0,3155	0,4055	0,5360

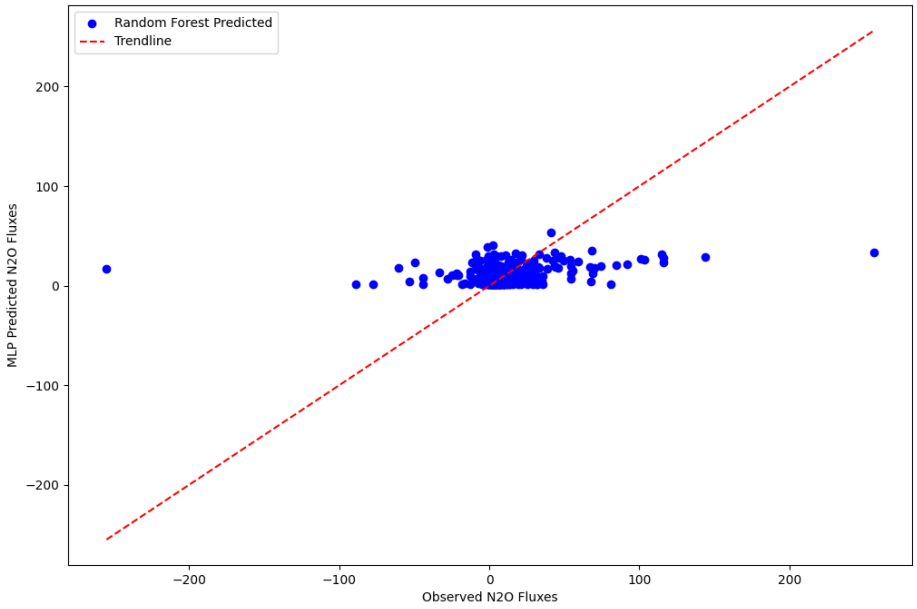

Figura 10 – Comparação entre valores observados e preditos de emissões de N_2O utilizando o modelo MLP.

De forma geral, os resultados obtidos indicam que o algoritmo *Random Forest* apresenta maior robustez para a modelagem de emissões de N_2O em ambientes agrícolas, es-

pecialmente em conjuntos de dados heterogêneos e desbalanceados. Esses achados estão em consonância com a literatura, que aponta o RF como uma abordagem eficaz para dados ambientais complexos, devido à sua capacidade de lidar com variáveis correlacionadas e capturar interações não lineares entre os preditores.

4.4 Importância de variáveis e efeitos parciais

A importância por permutação consiste em quantificar o aumento percentual do erro do modelo decorrente da permutação aleatória de uma variável individual, mantendo as demais constantes. Nesse procedimento, o erro quadrático médio da raiz (RMSE) é utilizado como métrica para avaliar de que forma a perturbação de uma variável específica influencia o erro médio das previsões de emissões de N_2O. Conforme ilustrado na Fig. 11, o maior incremento no erro esteve associado a variáveis diretamente relacionadas à disponibilidade de nitrogênio no solo, destacando-se o NH_4^+ como a variável mais influente (1,3%), seguido pelo NO_3^- (0,8%).

Embora as variáveis ligadas ao nitrogênio tenham apresentado maior relevância, parâmetros climáticos, como a umidade relativa média do ar (RH), e atributos físicos do solo, incluindo a fração de poros preenchidos por água (WFPS) e a temperatura do solo (TS), também figuraram entre as cinco variáveis mais importantes dentre as doze analisadas, evidenciando a natureza multifatorial do processo de emissão de N_2O.

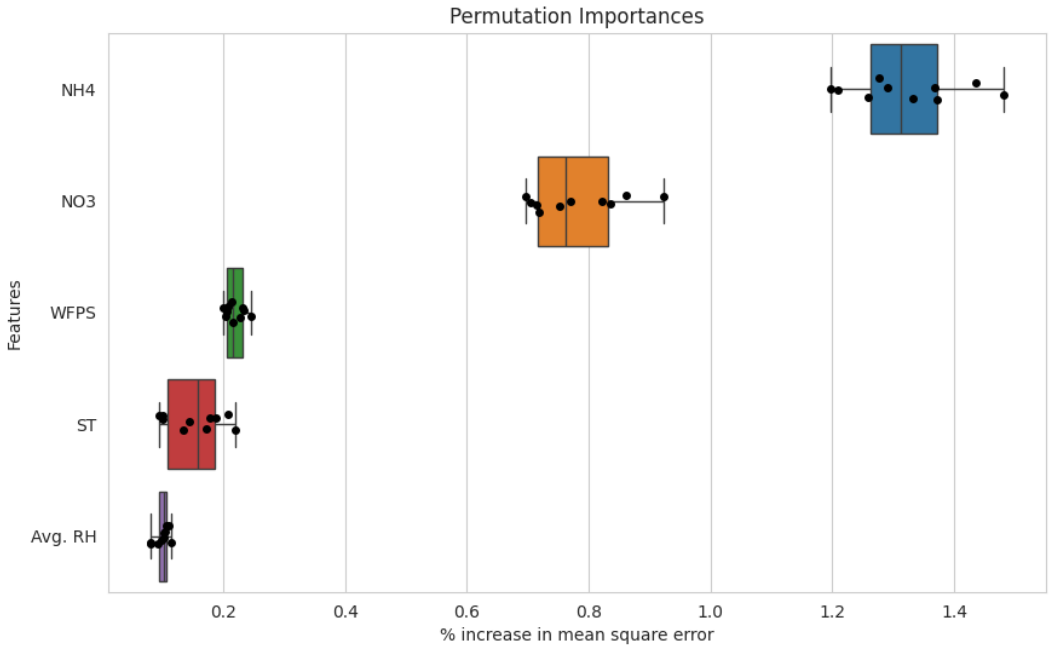

Figura 11 – Importância de variáveis por *Permutation Importance*.

A métrica de importância por permutação foi adotada neste estudo devido à sua

menor sensibilidade a distorções provocadas por multicolinearidade entre variáveis e por diferenças de escala, quando comparada a outros métodos de avaliação de importância, como a importância de Gini (Fig. 12). A importância de Gini, fundamentada nos critérios de divisão das árvores de decisão, tende a superestimar variáveis que apresentam maior número de categorias ou valores distintos. Em contraste, a importância por permutação avalia diretamente o impacto de cada variável sobre o desempenho preditivo do modelo, de forma independente do tipo de variável ou do algoritmo de aprendizado de máquina empregado.

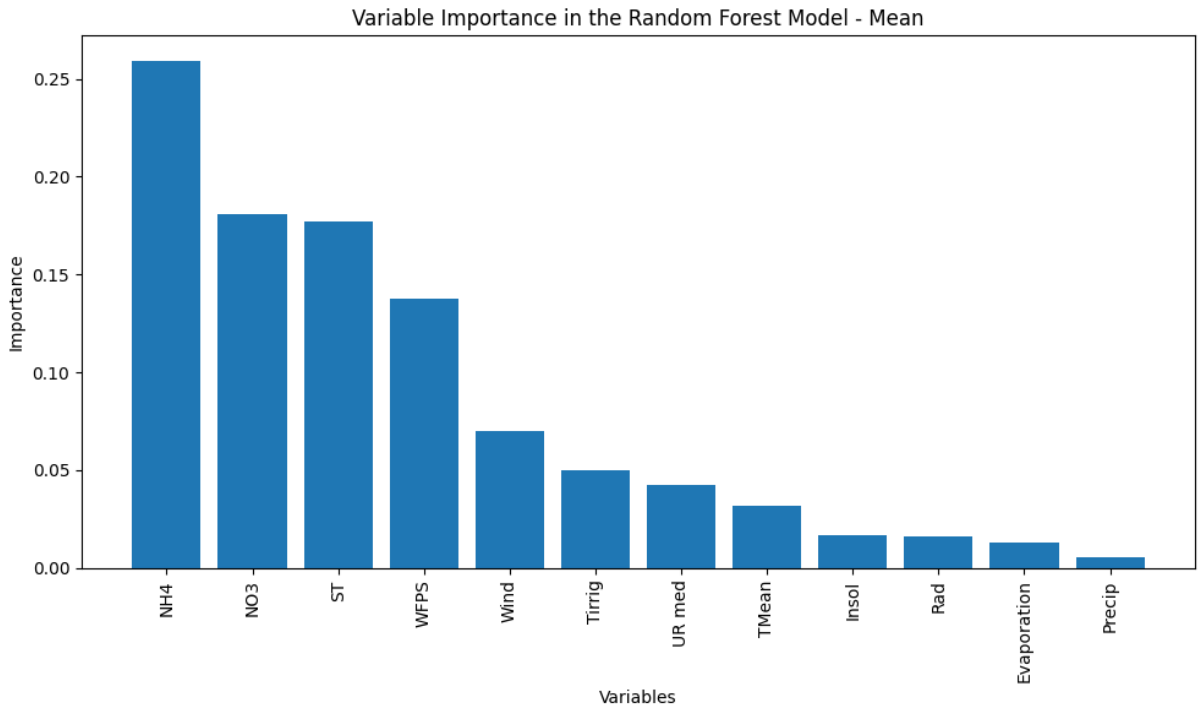

Figura 12 – Importância de variáveis por Random Forest (Gini).

A expectativa condicional individual (*Individual Conditional Expectation* – ICE) é uma ferramenta estatística empregada para analisar a influência de uma variável específica sobre a saída de um modelo preditivo. Para cada observação do conjunto de dados, o método calcula os valores previstos ao longo de uma faixa de variação da variável de interesse, mantendo constantes todas as demais variáveis. Dessa forma, o ICE permite avaliar a dependência parcial da resposta do modelo em relação a uma variável específica. As curvas individuais resultantes dessas predições são apresentadas na Fig. 13, evidenciando como a saída do modelo se altera em função das mudanças nas variáveis selecionadas.

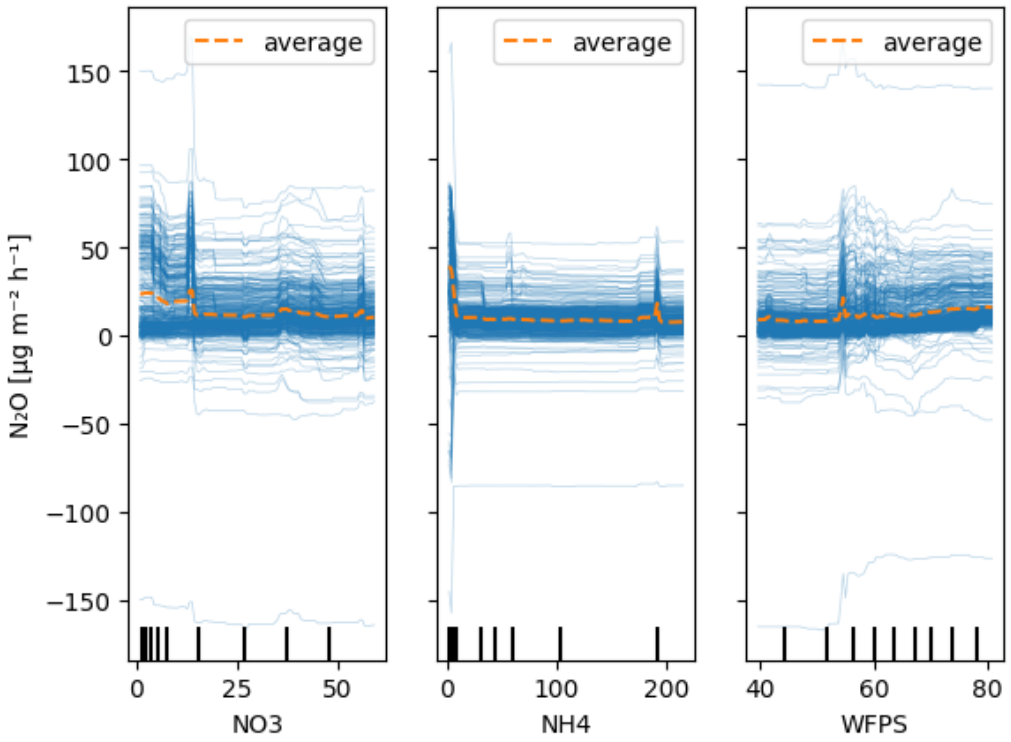

Figura 13 – ICE plots dos fluxos preditos de N_2O.

Na Fig. 13, os pontos coloridos representam a correlação das previsões associadas aos valores de NH_4^+, enquanto a linha preta indica a tendência média dos dados. Essa representação gráfica permite interpretar a variação da contribuição parcial de cada característica, além de servir como referência para a distribuição cromática dos valores e facilitar a visualização da relação entre as concentrações de NH_4^+ e as emissões previstas de N_2O.

Uma análise mais detalhada por meio do gráfico ICE correspondente revela uma elevada concentração de amostras próximas a valores nulos, além da ocorrência de picos acentuados de emissão nos valores próximos de 0 e 200, apontando esses intervalos como os mais representativos para essa variável. Para o NO_3^-, o gráfico ICE evidencia grande variabilidade da resposta do modelo para valores inferiores a 13, com recorrência de emissões entre 0 e 80 $\mu g\, m^{-2}\, h^{-1}$ de N_2O.

No caso do WFPS, o gráfico ICE apresenta uma tendência crescente das emissões de N_2O até aproximadamente 55%, faixa na qual os valores previstos situam-se entre 20 e 50 $\mu g\, m^{-2}\, h^{-1}$. A partir desse ponto, observa-se um comportamento de saturação da resposta do modelo, especialmente próximo a 80%. De modo geral, os gráficos ICE fornecem uma representação visual clara das relações entre variáveis individuais e as emissões de N_2O, destacando padrões específicos da resposta do modelo às variações nas condições ambientais e do solo.

As quatro variáveis consideradas mais relevantes foram analisadas de forma mais aprofundada quanto à sua contribuição para as previsões de emissões de N_2O (Fig. 14). A

ausência de acúmulo de NH_4^+ no solo sugere a ocorrência de reações de transformação do nitrogênio, resultando na conversão completa desse íon e, consequentemente, em variações pouco expressivas nos fluxos de N_2O. De forma geral, observou-se que os fluxos de N_2O tendem a ser mais elevados quando o WFPS ultrapassa o limiar de 55% e a temperatura do solo (TS) atinge valores próximos a 22 °C.

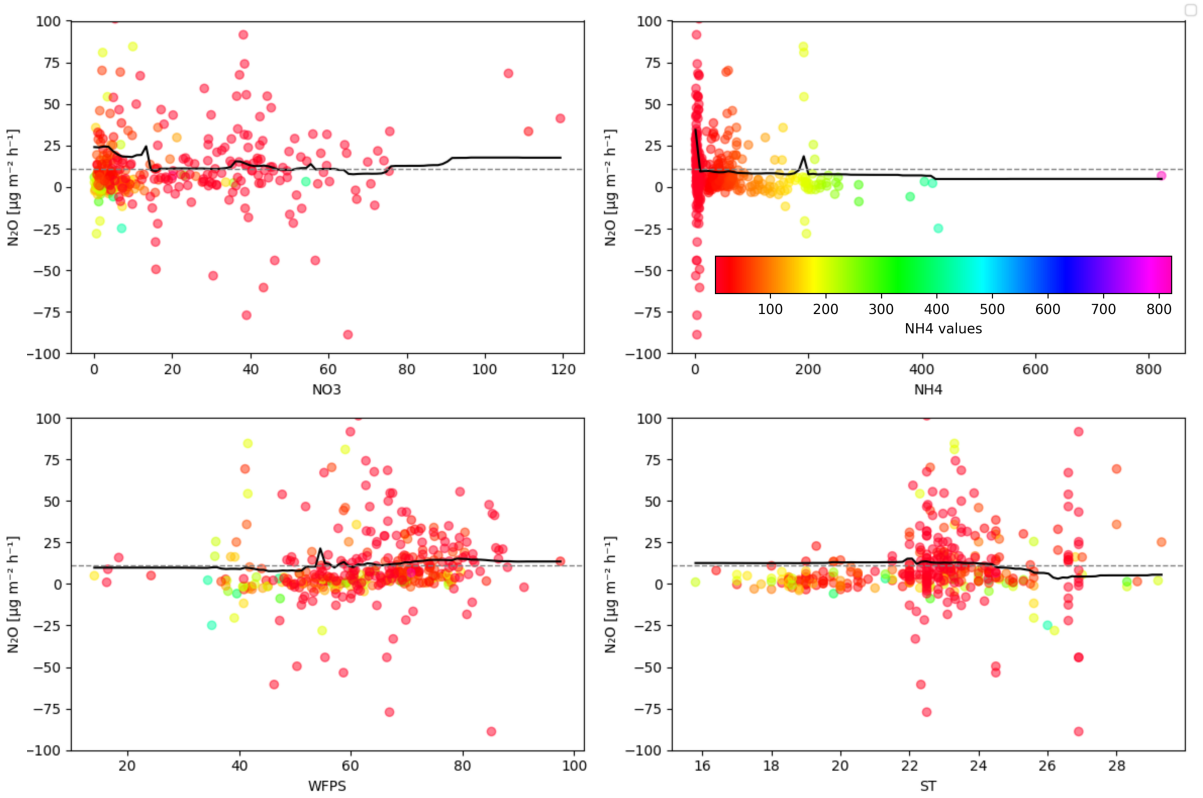

Figura 14 – Contribuições parciais das variáveis (NO_3^-, NH_4^+, WFPS, ST).

Os resultados obtidos neste estudo (Figs. 13 e 14) indicam que as variáveis NH_4^+ e NO_3^- apresentam maior poder explicativo sobre as emissões de N_2O quando comparadas à TS e ao WFPS. Embora todas essas variáveis atuem como fatores condicionantes das emissões, o ranqueamento de importância gerado pelo modelo reforça a interpretação de que NH_4^+ e NO_3^- constituem os principais elementos diretamente envolvidos na formação do N_2O, enquanto a TS e o WFPS desempenham o papel de condições ambientais que favorecem as transformações bioquímicas do nitrogênio no solo. Esse comportamento está em concordância com resultados reportados na literatura (Firestone; Davidson, 1989), que indicam que, sob condições de baixa disponibilidade de oxigênio, como ocorre em solos recentemente irrigados, há acúmulo de NH_4^+ devido à inibição de sua conversão em NO_3^-, não sendo observado, nesses casos, um aumento significativo nas emissões de N_2O.

De forma geral, a simulação realizada fornece subsídios relevantes para a compreensão de como variações em variáveis individuais influenciam as emissões de N_2O, conforme capturado pelo modelo. Os resultados evidenciam que a modelagem das emissões de N_2O

em sistemas de cana-de-açúcar por meio do algoritmo *Random Forest* apresenta vantagens em relação a esforços de modelagem anteriores realizados com o mesmo conjunto de dados.

4.5 Efeitos dos Regimes de Irrigação

Os diferentes regimes de irrigação apresentaram efeitos contrastantes sobre as emissões de N_2O. A irrigação de salvamento resultou em menores fluxos médios, mas com picos de emissão logo após a aplicação de água. Os regimes de reposição de 46% e 75% da ETc apresentaram maior emissão acumulada ao longo do ciclo da cultura, sendo o último o de maior impacto.

O regime de 46% da ETc mostrou-se mais equilibrado, conciliando manutenção da produtividade agrícola com menor intensificação das emissões em comparação ao regime de 75%. Esse resultado sugere que ajustes intermediários de irrigação podem representar alternativa viável para reduzir emissões sem comprometer a produção.

4.6 Comparação com Estudos Anteriores

Os resultados obtidos neste estudo corroboram achados de pesquisas anteriores que apontam forte influência do conteúdo de água no solo sobre as emissões de N_2O (Bonato *et al.*, 2026). Estudos realizados em outras regiões canavieiras também identificaram maior emissão em condições de alta umidade e disponibilidade de nitrogênio, bem como efeito intensificador da vinhaça.

A superioridade do RF em relação a outros algoritmos está em linha com aplicações recentes na modelagem de fluxos de gases de efeito estufa em solos agrícolas. Isso reforça a aplicabilidade desse método como ferramenta preditiva para apoiar decisões de manejo.

4.7 Implicações para o Manejo Sustentável

Os achados desta pesquisa têm implicações práticas relevantes para o manejo da cana-de-açúcar no Cerrado. A identificação de variáveis-chave permite direcionar estratégias de mitigação, como:

- Ajuste do regime de irrigação para níveis intermediários (46% da ETc), reduzindo o consumo desnecessário de água, ou seja, promovendo economia de água.
- Planejamento da aplicação de fertilizantes, priorizando momentos e quantidades que minimizem a sobreposição com altas taxas de irrigação.
- Monitoramento de variáveis convencionais (WFPS e NO_3^-) como indicadores de risco de intensificação das emissões.

Essas práticas podem contribuir para reduzir a pegada de carbono da cadeia produtiva da cana-de-açúcar, atendendo às demandas de sustentabilidade ambiental e competitividade no mercado internacional.

Essas práticas, fundamentadas no manejo de irrigação de precisão e no ajuste do período ideal de aplicação de fertilizantes, contribuem para reduzir a pegada de carbono da cadeia produtiva. Ao evitar o desperdício de insumos nitrogenados e otimizar o uso da água, reduz-se a emissão de óxido nitroso (N_2O) e o consumo energético do bombeamento, atendendo às demandas de sustentabilidade ambiental e competitividade no mercado internacional.

5 Conclusões

5.1 Síntese dos resultados

O presente estudo aplicou técnicas de aprendizado de máquina para a predição dos fluxos de N_2O em sistemas de cultivo de cana-de-açúcar submetidos a diferentes regimes de irrigação no bioma Cerrado. A integração de variáveis do solo, climáticas e de manejo de irrigação em modelos preditivos permitiu capturar relações complexas e não lineares que dificilmente seriam representadas por métodos estatísticos tradicionais. Entre os algoritmos avaliados, o modelo *Random Forest* apresentou desempenho superior, alcançando coeficiente de determinação (R^2) de 0,8736, além de baixos valores de erro absoluto médio (MAE) e erro quadrático médio (RMSE), o que evidencia sua elevada capacidade de identificação de tendências, embora o modelo apresente restrições na previsão acurada de magnitudes de fluxo em valores extremos.

A análise da importância das variáveis indicou que os principais fatores associados às emissões de N_2O estão diretamente ligados à disponibilidade de nitrogênio no solo e às condições físicas do ambiente. As concentrações de NH_4^+ e NO_3^- foram identificadas como as variáveis mais influentes, seguidas pela temperatura do solo (ST) e pela fração de poros preenchidos por água (WFPS). Esses resultados reforçam o papel central dos processos de transformação do nitrogênio na formação do N_2O, enquanto ST e WFPS atuam como condicionantes ambientais que modulam a intensidade dessas reações bioquímicas.

Os regimes de irrigação exerceram influência significativa sobre os fluxos de N_2O. A irrigação correspondente a 75% da evapotranspiração da cultura promoveu as maiores emissões, indicando que elevados níveis de umidade do solo favorecem condições propícias à produção de N_2O. Por outro lado, o regime de reposição de 46% da ETc apresentou um equilíbrio mais adequado entre a manutenção da produtividade agrícola e a mitigação das emissões. Observou-se ainda que a aplicação de vinhaça intensificou os fluxos de N_2O, especialmente quando associada a regimes de maior irrigação, evidenciando a necessidade de estratégias integradas de manejo da fertilização e da irrigação.

5.2 Contribuições

Os resultados deste trabalho contribuem para o avanço do conhecimento científico ao demonstrar a aplicabilidade e a robustez de modelos de aprendizado de máquina na análise de emissões de N_2O em sistemas agrícolas tropicais. A abordagem adotada permitiu identificar interações não lineares entre variáveis edafoclimáticas e práticas de manejo,

ampliando a compreensão dos fatores que controlam as emissões desse gás de efeito estufa em áreas de cana-de-açúcar no Cerrado.

Do ponto de vista prático, os achados fornecem subsídios relevantes para a definição de estratégias de manejo mais sustentáveis, especialmente no que se refere à escolha de regimes de irrigação e ao uso da vinhaça. A identificação das variáveis mais influentes sobre as emissões possibilita direcionar ações de mitigação de forma mais eficiente, conciliando produtividade agrícola e redução de impactos ambientais. Adicionalmente, a disponibilização dos dados e do código utilizado em repositório aberto reforça o caráter reprodutível do estudo e favorece sua utilização e ampliação por outros pesquisadores.

5.3 Limitações e trabalhos futuros

Apesar dos avanços obtidos, algumas limitações devem ser consideradas. O conjunto de dados utilizado apresenta restrições em termos de série temporal, o que pode ter limitado o desempenho de algoritmos mais complexos, como redes neurais profundas. Além disso, o experimento foi conduzido em uma única área representativa do bioma Cerrado, o que impõe cautela na extrapolação dos resultados para outras regiões, condições de solo ou contextos climáticos distintos. A ausência de variáveis de longo prazo, como a variabilidade interanual do clima e diferentes ciclos de cultivo da cana-de-açúcar, também constitui uma limitação do estudo.

Nesse sentido, pesquisas futuras devem priorizar a ampliação da base de dados, incorporando medições em diferentes safras, solos e condições climáticas. A exploração de algoritmos híbridos e técnicas de *ensemble learning*, bem como a integração de sensores em tempo real para alimentar modelos preditivos dinâmicos, pode contribuir para o aprimoramento das previsões. A incorporação de modelos baseados em processos biogeoquímicos ao arcabouço de aprendizado de máquina também representa uma perspectiva promissora, ao permitir a inclusão explícita de mecanismos microbianos e indicadores de saúde do solo.

5.4 Considerações finais

Este trabalho evidencia o potencial do aprendizado de máquina como ferramenta de apoio à agricultura sustentável no Cerrado brasileiro. A capacidade dos modelos em representar relações não lineares e em simular cenários a partir da variação dos parâmetros de entrada amplia as possibilidades de uso dessa abordagem tanto para análises retrospectivas quanto para projeções futuras.

Ao identificar variáveis-chave e propor estratégias de manejo de irrigação mais eficientes, a pesquisa contribui para a mitigação das emissões de gases de efeito estufa e para o desenvolvimento de sistemas agrícolas mais resilientes. Dessa forma, os resultados apre-

sentados oferecem subsídios relevantes para políticas públicas, planejamento agrícola e iniciativas voltadas à adaptação e mitigação às mudanças climáticas.

Em última análise, este trabalho não se encerra em suas conclusões atuais; ele estabelece a base fundamental para um monitoramento contínuo e estratégico. Projeta-se que a manutenção rigorosa dessas diretrizes garanta não apenas a estabilidade dos resultados alcançados, mas também a robustez necessária para que o projeto se adapte com segurança às oscilações e desafios do cenário externo nos próximos anos.

Referências

ALPAYDIN, E. **Introduction to Machine Learning**. Cambridge, MA: MIT Press, 2004. Citado nas pp. 26, 27.

ARNOLD, J. G.; SRINIVASAN, R.; MUTTIAH, R. S.; WILLIAMS, J. R. LARGE AREA HYDROLOGIC MODELING AND ASSESSMENT PART I: MODEL DEVELOPMENT. **JAWRA Journal of the American Water Resources Association**, v. 34, n. 1, p. 73–89, 1998. DOI: https://doi.org/10.1111/j.1752-1688.1998.tb05961.x. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1752-1688.1998.tb05961.x. Disponível em: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1752-1688.1998.tb05961.x. Citado na p. 18.

BELGIU, M.; DRĂGUŢ, L. Random forest in remote sensing: a review of applications and future directions. **ISPRS Journal of Photogrammetry and Remote Sensing**, v. 114, p. 24–31, 2016. Citado na p. 26.

BONATO, R. T.; BORGES, L. d. A. B.; CARVALHO, A. M.; OLIVEIRA, A. D.; SOUSA, T. R.; RAMOS, M. L. G.; RIBEIRO JUNIOR, W. Q.; MARCHÃO, R. L.; SILVA, F. A. M.; BORGES, D. L. Predicting nitrous oxide emissions from soil planted to sugarcane under various irrigation regimes using machine learning models. **Frontiers of Agricultural Science and Engineering**, v. 13, p. 25663, 2026. ISSN 2097-7654. DOI: 10.15302/J-FASE-2025663. Disponível em: https://journal.hep.com.cn/fase/EN/10.15302/J-FASE-2025663. Citado na p. 39.

BRASIL; MINISTÉRIO DO DESENVOLVIMENTO, INDÚSTRIA, COMÉRCIO E SERVIÇOS. **Nova Indústria Brasil: política industrial para o desenvolvimento produtivo, inovador e sustentável**. Brasília, 2024. Citado na p. 13.

BREIMAN, L. Random Forests. **Machine Learning**, v. 45, n. 1, p. 5–32, 2001. Citado na p. 25.

BUTTERBACH-BAHL, K.; DANNENMANN, M. Denitrification and associated soil N2O emissions due to agricultural activities in a changing climate. **Current Opinion in Environmental Sustainability**, v. 3, n. 5, p. 389–395, 2011. Carbon and nitrogen cycles. ISSN 1877-3435. DOI: https://doi.org/10.1016/j.cosust.2011.08.004. Disponível em: https://www.sciencedirect.com/science/article/pii/S1877343511000698. Citado na p. 12.

CARVALHO, A. M. de; OLIVEIRA, A. D. de; COSER, T. R.; SOUSA, T. R. de; LIMA, C. A. de; RAMOS, M. L. G.; MALAQUIAS, J. V.; ARAUJO GONÇALVES, A. D. M. de; JÚNIOR, W. Q. R. N2O emissions from sugarcane fields under contrasting watering regimes in the Brazilian Savannah. **Environmental Technology & Innovation**, Elsevier, v. 22, p. 101470, 2021. Citado nas pp. 18, 20–24.

FAN, G. F.; ZHANG, L. Z.; YU, M.; HONG, W. C.; DONG, S. Q. Applications of random forest in multivariable response surface for short-term load forecasting. **International Journal of Electrical Power & Energy Systems**, v. 139, p. 108073, 2022. Citado na p. 26.

FIRESTONE, M. K.; DAVIDSON, E. A. Microbiological basis of NO and N2O production and consumption in soil. *In:* ANDREAE, M. O.; SCHIMEL, D. S. (ed.). **Exchange of Trace Gases Between Terrestrial Ecosystems and the Atmosphere**. New York: John Wiley & Sons, 1989. p. 7–21. Citado na p. 38.

FISCHER, A. How to determine the unique contributions of input-variables to the nonlinear regression function of a multilayer perceptron. **Ecological Modelling**, v. 309–310, p. 60–63, 2015. Citado na p. 26.

HAMOUD, Y. A.; SHAGHALEH, H. *et al.* Predicting nitrous oxide emissions from soil planted to sugarcane under various irrigation regimes using machine learning models. **Frontiers of Agricultural Science and Engineering**, 2025. Verificar volume/página final, publicação recente. DOI: VerificarDOInapublicaÃğÃčofinal. Citado na p. 13.

INTERGOVERNMENTAL PANEL ON CLIMATE CHANGE (IPCC). **IPCC Guidelines for National Greenhouse Gas Inventories**. Edição: H. S. Eggleston, L. Buendia, K. Miwa, T. Ngara e K. Tanabe. Hayama: Institute for Global Environmental Strategies (IGES), 2006. Citado na p. 22.

ISLAM, A. R. M. T.; TALUKDAR, S.; MAHATO, S.; KUNDU, S.; EIBEK, K. U.; PHAM, Q. B.; KURIQI, A.; LINH, N. T. T. Flood susceptibility modelling using advanced ensemble machine learning models. **Geoscience Frontiers**, v. 12, n. 3, p. 101075, 2021. Citado na p. 26.

LEITE, E. A.; SILVA, E. H. F. M. da; FATTORI JÚNIOR, I. M.; MARIN, F. R. Assessing climate change impacts on sugarcane yield, crop water productivity, and nitrous oxide emissions across Brazil's bioenergy using the CSM-SAMUCA-sugarcane model. **Agricultural Systems**, Elsevier, v. 231, p. 104118, 2026. Este é o estudo sobre cenários futuros e modelagem climática anteriormente atribuído ao grupo da USP/Esalq. DOI: 10.1016/j.agsy.2025.104118. Citado na p. 13.

LI, Y.; CHEN, D.; ZHANG, Y.; EDIS, R.; DING, H. Comparison of three modeling approaches for simulating denitrification and nitrous oxide emissions from loam-textured arable soils. **Global Biogeochemical Cycles**, v. 19, n. 3, 2005. DOI: https://doi.org/10.1029/2004GB002392. eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2004GB002392. Disponível em: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2004GB002392. Citado na p. 18.

P.R. DA SILVA, V. de; DA SILVA, B. B.; ALBUQUERQUE, W. G.; BORGES, C. J.; DE SOUSA, I. F.; NETO, J. D. Crop coefficient, water requirements, yield and water use efficiency of sugarcane growth in Brazil. **Agricultural Water Management**, v. 128, p. 102–109,

2013. ISSN 0378-3774. DOI: https://doi.org/10.1016/j.agwat.2013.06.007. Citado na p. 21.

PARTON, W. J.; HOLLAND, E. A.; DEL GROSSO, S. J.; HARTMAN, M. D.; MARTIN, R. E.; MOSIER, A. R.; OJIMA, D. S.; SCHIMEL, D. S. Generalized model for NO x and N2O emissions from soils. **Journal of Geophysical Research: Atmospheres**, v. 106, n. D15, p. 17403–17419, 2001. DOI: https://doi.org/10.1029/2001JD900101. eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2001JD900101. Disponível em: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2001JD900101. Citado na p. 18.

PORTMANN, R. W.; DANIEL, J. S.; RAVISHANKARA, A. R. Stratospheric ozone depletion due to nitrous oxide: influences of other gases. **Philosophical Transactions of the Royal Society B: Biological Sciences**, v. 367, n. 1593, p. 1256–1264, maio 2012. ISSN 0962-8436. DOI: 10.1098/rstb.2011.0377. eprint: https://royalsocietypublishing.org/rstb/article-pdf/367/1593/1256/102079/rstb.2011.0377.pdf. Disponível em: https://doi.org/10.1098/rstb.2011.0377. Citado na p. 12.

SAHA, D.; BASSO, B.; ROBERTSON, G. P. Machine learning improves predictions of agricultural nitrous oxide (N2O) emissions from intensively managed cropping systems. **Environmental Research Letters**, IOP Publishing, v. 16, n. 2, p. 024004, 2021. Citado na p. 18.

SCARPARE, F. V.; HERNANDES, T. A. D.; RUIZ-CORRÊA, S. T.; PICOLI, M. C. A.; SCANLON, B. R.; CHAGAS, M. F.; DUFT, D. G.; FÁTIMA CARDOSO, T. de. Sugarcane land use and water resources assessment in the expansion area in Brazil. **Journal of cleaner production**, Elsevier, v. 133, p. 1318–1327, 2016. Citado na p. 13.

SILVA, A. C. *et al.* Simulating soil water dynamics, soybean yield and nitrous oxide emission under conventional and no-tillage systems in the Brazilian Cerrado using STICS model. **Agriculture, Ecosystems & Environment**, Elsevier, v. 361, p. 108842, 2024. DOI: 10.1016/j.agee.2023.108842. Citado na p. 18.

TIAN, H.; XU, R.; CANADELL, J. G.; THOMPSON, R. L.; WINIWARTER, W.; SUNTHARALINGAM, P.; DAVIDSON, E. A.; CIAIS, P.; JACKSON, R. B.; JANSSENS-MAENHOUT, G. *et al.* A comprehensive quantification of global nitrous oxide sources and sinks. **Nature**, Nature Publishing Group, v. 586, n. 7828, p. 248–256, 2020. Citado na p. 12.

YIN, Y.; WANG, Z.; TIAN, X.; WANG, Y.; CONG, J.; CUI, Z. Evaluation of variation in background nitrous oxide emissions: A new global synthesis integrating the impacts of climate, soil, and management conditions. **Global Change Biology**, Wiley Online Library, v. 28, n. 2, p. 480–492, 2022. Citado na p. 12.

Apêndices

APÊNDICE A – Guia Operacional: Processamento, Modelagem e Visualização de Dados

Este apêndice apresenta um guia detalhado para a operação do código disponibilizado no GitHub e desenvolvido em Python para análise de dados agrícolas e predição de fluxos de N_2O. A implementação utiliza a biblioteca (**scikit-learn**), ferramenta utilizada para algoritmos de aprendizado de máquina em ambiente científico.

A.1 Preparação e Estrutura do Código

O código deve ser executado preferencialmente no ambiente **Google Colab**. Para iniciar, o usuário deve carregar o arquivo de dados (`Dados_RF.csv`) através da célula inicial de *upload*. O código está organizado nos seguintes blocos lógicos:

- **Carga e Limpeza:** Células iniciais onde os dados são importados e as variáveis (como *ST* e *WFPS*) são renomeadas.
- **Modelagem (RF e MLP):** Blocos identificados pelas funções `RandomForestRegressor` e `MLPRegressor`.
- **Análise de Importância:** Seção final contendo cálculos de *Gini* e *Permutation Importance*.
- **Visualização Avançada:** Células destinadas à geração de *ICE plots*.

A.2 Execução dos Modelos Preditivos

O sistema permite a comparação entre dois modelos de aprendizado de máquina:

A.2.1 Random Forest (RF)

O modelo de Florestas Aleatórias é robusto para lidar com multiplas variáveis .

1. Localize a célula de treinamento do RF e execute-a.

2. O código retornará automaticamente as métricas de desempenho por bloco (MAE, MSE e R^2). O valor de R^2 indica a precisão do modelo (quanto mais próximo de 1.0, melhor).

A.2.2 Multi-Layer Perceptron (MLP)

Trata-se de uma rede neural artificial capaz de captar relações não-lineares complexas.

1. Execute a célula da MLP após o processamento dos dados.
2. Observe que este modelo exige maior capacidade computacional, podendo levar alguns minutos a mais que o RF para concluir as iterações.

A.3 Geração de Imagens e Gráficos de Saída

Para a documentação dos resultados, o código gera três tipos fundamentais de visualizações:

1. **Importância de Variáveis (Gini):** Gera um gráfico de barras mostrando o impacto direto de cada variável (ex: Nitrato, Umidade) na decisão das árvores do RF. Localize o comando `feature_importances_`.
2. **Permutation Importance:** Uma análise mais rigorosa que valida a importância das variáveis através da permutação dos dados. Execute a célula que invoca a função `permutation_importance` da biblioteca (**scikit-learn**).
3. **ICE Plots (Individual Conditional Expectation):** Estes gráficos mostram como as predições de N_2O variam conforme a mudança de uma única variável de entrada. Procure pela função `PartialDependenceDisplay` para gerar estas curvas.

A.4 Citações de Ferramentas Utilizadas

A principal biblioteca utilizada para a construção dos modelos e validação estatística é a *scikit-learn*:

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

A.5 Dicas de Resolução de Problemas

Caso o código apresente erros de execução:

- Verifique se o nome do arquivo carregado é exatamente `Dados_RF.csv`.
- Certifique-se de executar as células em ordem sequencial (do topo para baixo).
- Se os gráficos não aparecerem, verifique se a biblioteca `matplotlib` foi importada corretamente na primeira célula.