UNIVERSIDADE FEDERAL DE SÃO CARLOS (UFSCAR)
CENTRO DE CIÊNCIAS EXATAS E DE TECNOLOGIA
GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

GABRIEL PANDOLFI CORREA DO SANTOS

QUEIMADAS E FATORES ASSOCIADOS A PRODUTIVIDADE PRIMÁRIA BRUTA NO CERRADO: UMA ABORDAGEM DE *MACHINE LEARNING* EXPLICÁVEL

TRABALHO DE CONCLUSÃO DE CURSO

SÃO CARLOS (SP)
2025

QUEIMADAS E FATORES ASSOCIADOS A PRODUTIVIDADE PRIMÁRIA BRUTA NO CERRADO: UMA ABORDAGEM DE *MACHINE LEARNING* EXPLICÁVEL

Trabalho de Conclusão de curso apresentado ao Curso de Ciência da Computação como requisito para graduação em Ciência da Computação, Universidade de São Carlos.

Aprovado em: 24 de fevereiro de 2025.

BANCA EXAMINADORA

_____

Profª. Orientadora: Drª. Heloisa de Arruda Camargo
Universidade Federal de São Carlos (UFSCAR)

_____

Prof. Dr. Alan Demétrius Baria Valejo
Universidade Federal de São Carlos (UFSCAR)

_____

Prof. Dr. Alexandre Luís Magalhães Levada
Universidade Federal de São Carlos (UFSCAR)

GABRIEL PANDOLFI CORREA DO SANTOS

QUEIMADAS E FATORES ASSOCIADOS PRODUTIVIDADE PRIMÁRIA BRUTA NO CERRADO: UMA ABORDAGEM DE *MACHINE LEARNING* EXPLICÁVEL

Trabalho de Conclusão de curso apresentado ao Curso de Ciência da Computação como requisito para graduação em Ciência da Computação, Universidade de São Carlos.

Orientadora: Profª. Drª. Heloisa de Arruda Camargo

SÃO CARLOS (SP)
2025

Ficha Catalográfica

Ficha catalográfica elaborada automaticamente de acordo com os dados fornecidos pelo autor: Pandolfi, Gabriel Correa dos Santos.

Queimadas e fatores associados à Produtividade Primária Bruta do Cerrado: Uma Abordagem de *Machine Learning*/Pandolfi, Gabriel Pandolfi Correa dos Santos. São Carlos (SP), fev. 2025. 78 f. 30 cm.
Modo de acesso:
Disponível em:
Inclui bibliografia: p.65-68 Inclui ilustrações.

Orientadora: Profª. Drª. Heloisa de Arruda Camargo.
Trabalho de Conclusão de curso (TCC) - Graduação em Ciência da Computação, Universidade Federal de São Carlos, São Carlos, 2025.

1. Ciência da Computação: I. Focos de Incêndios; II. Produtividade Primária Bruta (PPB); III. Cerrado; IV. Aprendizado de Máquina; V. *Shapley Additive Explanations* (SHAP). Título: **Queimadas e fatores associados à Produtividade Primária Bruta do Cerrado**: Uma Abordagem de *Machine Learning*.

LISTA DE ABREVIATURAS E SIGLAS

AdaBoost	*Adaptive Boosting*
AM	Amazonas
BA	Bahia
BEPS	*Base Erosion Profit Shifting*
CART	*Classification and Regression Trees*
CHAID	Qui-quadrado
ID3	Ganho de Informação
CLI	Consentimento Livre Informado
CO^2	Carbono Atmosférico
DARPA	*Defense Advanced Research Projects Agency*
DF	Distrito Federal
EC-LUE	Modelo de Eficiência de Uso de Luz
EOS	*Earth Observing System*
ESE	*Earth Science Enterprises*
ET	*Extra Trees*
FPAR	*Fraction of Photosynthetically Active Radiation* do *MOD15A2H* 6.1
GEE	Gases de Efeito Estufa
GO	Goiás
GPP	*Growth Primary Production*
IA	Inteligência Artificial
IBGE	Instituto Brasileiro de Geografia e Estatística
INMET	Instituto Nacional de Meteorologia
INPE	Instituto Nacional de Pesquisas Espaciais
LIME	*Local Interpretable Model-Agnostic Explanations*
LAI	*Leaf Area Index (LAI)*
LST&E	*Land Surface Temperature and Emissivity*
MA	Maranhão
MAE	Erro Absoluto Médio
MAPE	Erro Percentual Absoluto Médio
MG	Minas Gerais
ML	*Machine Learning*

MMA	Ministério do Meio Ambiente e Mudança do Clima
MODIS	*Moderate-resolution Imaging Spectroradiometer*
MOD11A2	*Land Surface Temperature/Emissivity 8-Day* 6.1
MS	Mato Grosso do Sul
MSE	Erro Quadrático Médio
MT	Mato Grosso
NaN	*Not a Number*
NASA	*National Aeronautics and Space Administration*
NEE	Produtividade Ecossistêmica Líquida
NPP	Produtividade primária líquida média (PPLM) de longo prazo
ONG	Organização Não-governamental
PCD	Pessoas com Deficiência
PI	Piauí
PIB	Produto Interno Bruto
PPL	Produtividade Primária Líquida
PPLM	Produtividade Primária Líquida Média
QV	Qualidade de Vida
R	Respiração
R^2	Coeficiente de Determinação
RF	*Random Forest*
RMSE	Erro Quadrático Médio
RMSLE	Raiz do Erro Logarítmico Quadrático
RO	Rondônia
SHAP	*Shapley Additive Explanations*
SP	São Paulo
TO	Tocantins
UFSCAR	Universidade Federal de São Carlos
VDD	*Déficit* de pressão de vapor atmosférico
XAI	Inteligência Artificial Explicável
XGBoost	Modelo *Extreme Gradient Boosting*

LISTA DE FIGURAS/GRÁFICOS

RESUMO

Introdução: A *Growth Primary Production* (GPP) é um dos componentes fundamentais para o planeta. Se relaciona com clima, vegetação, agronegócio, ciclo da água, saúde, bem como incêndios e queimadas, e consequentes mudanças climáticas, etc., sendo também a base da cadeia alimentar. **Objetivo:** Analisar as variáveis mais influentes da GPP, por meio de técnicas de Inteligência Artificial Explicável (XAI) e do *SHapley Additive exPlanations* (SHAP) no Cerrado brasileiro. **Metodologia**: Dez estações meteorológicas de três estados (Goiás, Mato Grosso e Mato Grosso do Sul) e do DF, compuseram a amostra. Os dados relativos às variáveis do Tempo, Clima, Vegetação e Incêndios, da série histórica de 2003-2020 foram obtidos do sensor remoto *Moderate Resolution Imaging Spectroradiometer* (MODIS)/*National Aeronautics and Space Administration* (NASA), do Instituto Nacional de Meteorologia (INMET) e do Programa Queimadas do Instituto Nacional de Pesquisas Espaciais (INPE). Foi realizada uma análise exploratória, pré-processamento, modelagem e inferência dos dados. As análises foram feitas por meio de Modelos Tipo Comitê (*Ensemble*), Árvores Aleatórias (*Random Forest*-RF), Árvores Extremamente Aleatórias (Extra Trees-ET) e *Adaptive Boosting* (Adaboost). **Resultados**: Os modelos que apresentaram melhor performance em relação à GPP foram ET, RF e *AdaBoost*. As variáveis mais importantes (*feature importance*) foram a Temperatura da superfície e Incêndios. **Conclusão**: O SHAP permitiu superar as limitações dos modelos preditivos tradicionais, favorecendo uma análise mais profunda dos fatores de influência da GPP. Variáveis como focos de incêndio e temperatura impactaram negativamente, ao contrário da vegetação que influenciou positivamente para a sustentabilidade do ecossistema do Cerrado brasileiro. Propõe-se expansão futura para incluir mais variáveis e modelos, aprimorando as análises de sustentabilidade ambiental.

Palavras chave: Focos de Incêndios; Produtividade Primária Bruta (PPB); Cerrado; Aprendizado de Máquina; *Shapley Additive Explanations* (SHAP)

ABSTRACT

Introduction: Primary Growth Production (GPP) is one of the fundamental components for the planet. It is related to climate, vegetation, agribusiness, water cycle, health, as well as fires and burning, and consequent climate changes, etc., and is also the base of the food chain. **Objective**: To analyze the most influential variables of Growth Primary Production (GPP), through Explainable Artificial Intelligence (XAI) and techniques and SHapley Additive ExPlanations (SHAP) in the Brazilian Cerrado. **Methodology**: Eleven meteorological stations from three states (Goiás, Mato Grosso and Mato Grosso do Sul) and the Federal District, composed the sample. The data related to the variables Weather, Climate and Vegetation and Fire outbreaks, of the 2003-2020 time series were obtained from the Moderate Resolution Imaging Spectroradiometer (MODIS)/*National Aeronautics and Space Administration* (NASA) remote sensor, from the National Institute of Meteorology (INMET) and from the Queimadas Program of the National Institute for Space Research (INPE). An exploratory analysis, pre-processing, modeling and inference of the data was carried out. The analyses were made using Ensemble Models, Random Forest (RF), Extra Trees (ET) and Adaptive Boosting. **Results**: The models that presented the best performance in relation to GPP were ET, RF and AdaBoost. The most important variables (feature importance) were Surface Temperature and Fires. **Conclusion**: SHAP made it possible to overcome the limitations of traditional predictive models, favoring a deeper analysis of the factors influencing GPP. Variables such as fire outbreaks and temperature had a negative impact, unlike vegetation, which had a positive impact on the sustainability of the Brazilian Cerrado ecosystem. Future expansion is proposed to include more variables and models, improving environmental sustainability analyses.

Key Words: Fire outbreaks; Growth Primary Production; Bushland; Machine Learning; Shapley Additive Explanations (SHAP).

SUMÁRIO

1. INTRODUÇÃO

No planeta Terra, há uma diversidade de biomas e *habitats* que abrigam incríveis criaturas que vivenciam uma realidade aparentemente sem ligação nenhuma com os seres humanos. Por vezes, parece até que cada ser vivo cuida de si, sem influenciar ou ser influenciado.

Contudo, isso é apenas uma falsa impressão, e tais relações provocam no ambiente global constante transformação. A fotossíntese, por exemplo, é, sem dúvida, um dos processos biológicos mais fundamentais na Terra, pois serve de alicerce para quase todas as formas de vida devido à sua capacidade de transformar energia, que é armazenada e é utilizada pelas plantas para crescer e se desenvolver (Pei *et al*., 2022).

No processo da fotossíntese não só sustenta a base da cadeia alimentar, mas também desempenha um papel crucial no ciclo do carbono, influenciando diretamente as trocas de CO^2 entre os ecossistemas terrestres e a atmosfera (Pei *et al*., 2022).

Ocorre que a exploração do ambiente, o estilo de vida e comportamento humanos colocam em risco esta relação. Desmatamento, mudanças climáticas, aquecimento global, agronegócio, poluição e gases nocivos estão associados aos riscos ambientais e à vida; e, isso não é diferente no Brasil e no cerrado brasileiro (De Santana, Delgado e Schiavetti, 2020).

O Brasil é um dos principais produtores e exportadores de soja e algodão, bem como de carne e grãos, proveniente principalmente de pastagens e áreas plantadas. O Cerrado contém a maior área de terras agrícolas e pecuárias do país (44%); produz 40% da carne, 84% do algodão, 60% da soja e 44% do milho. Explorado pelo agronegócio, o Cerrado é responsável por 23% do Produto Interno Bruto (PIB), colocando o Brasil no 9º lugar entre os maiores PIBs do mundo (IBGE, 2023); e, isso tem um custo ambiental.

Ao estudarem e proporem um Perfil do Ecossistema do Cerrado brasileiro Sawyer *et al*. (2017) ressaltaram que o Cerrado tem importância histórica, econômica, política, social e também para o ecossistema global. Para atingir tal

posição, todos os anos, há queima do bioma florestal - uma prática recorrente -, conhecida como queimadas, que acontece no período entre os meses de junho e novembro, nos quais há aumento de registro dos focos de calor conforme o Instituto Nacional de Pesquisas Espaciais (INPE) (Brasil, 2020), motivo de críticas de ambientalistas e Organizações Não-Governamentais (ONGs).

Ressalta-se que os incêndios e queimadas têm ocorrido em áreas de vegetação primária com alta diversidade, principalmente nos biomas Amazônia (67%), Cerrado (53%) e Pantanal (89%), e vêm aumentando. Dados de 1985-2023 do MapBiomas[1] revelam que o Brasil perdeu 15% das florestas naturais e os biomas com maior perda foram a Amazônia (13%) e o Cerrado (27%).

Seus efeitos afetam desde a produção de energia até a agricultura e o transporte de grãos, sem contar que a crise do clima está relacionada à insegurança alimentar. Secas prolongadas e padrões de chuva alterados no mundo, levam a produção de alimentos ao risco, com consequentes crises humanitárias e migrações.

Como se não bastasse, não se pode negar que tais fenômenos e alterações comprometem o ecossistema global e a biodiversidade, pois contribuem para a emissão de poluentes atmosféricos e podem interferir também com a saúde e Qualidade de Vida (QV) das pessoas, em especial dos mais vulneráveis como crianças, idosos e pessoas com deficiência (PcDs) (Brasil, 2020).

Tais questões emergem como problemas e desafios frequentes no país e no bioma do Cerrado. Fatores climáticos e do solo somados a focos de incêndio e às áreas afetadas pelas queimadas tendem a diminuir a *Growth Primary Production* (GPP) ou Produtividade Primária Bruta (PPB), uma parte essencial do ciclo da vida na terra e uma variável importante nos estudos do ciclo global do carbono, uma vez que define a taxa de extração de carbono atmosférico dos ecossistemas terrestres (Pei *et al.,* 2022; Danelichen *et al*., 2015).

A GPP é definida como a quantidade de carbono fixada pelos ecossistemas terrestres por fotossíntese, e é considerada a base para produção de vegetal. É um

[1] O MapBiomas é uma rede colaborativa formada por ONGs, universidades, laboratórios e *startups* que realiza o mapeamento anual da cobertura e uso da terra, além do monitoramento mensal da superfície de água e das cicatrizes de fogo com dados desde 1985. Utiliza processamento distribuído e automatizado de dados, em parceria com o *Google Earth Engine*. É uma plataforma aberta, escalável e projetada para ser aplicada em diferentes países e contextos, que busca a conservação e o manejo sustentável dos recursos naturais, como forma de combate às mudanças climáticas.

mecanismo responsável pelo fluxo global de carbono e se associa a diversas atividades e acontecimentos no ecossistema, compõem o principal processo que controla as trocas de CO^2 entre os ecossistemas terrestres e a atmosfera e tornou-se um importante parâmetro biofísico e pode auxiliar na compreensão e na dinâmica dos fluxos de carbono (Pei *et al*., 2022; Prăvălie *et al*., 2023), bem como no ciclo vital.

Analisar a GPP é importante para o conhecimento e tomada de decisão e preservação da vida. Para a atenuar a diminuição da GPP é preciso reduzir, se não mitigar, a degradação ambiental e a influência humana sobre o planeta, por meio de uma melhor gestão ambiental. Nos últimos anos, uma ampla gama de metodologias, incluindo modelos climáticos, métodos estatísticos são empregados para aumentar a precisão das estimativas da GPP a partir de perspectivas espaciais e temporais (Kamel, 2015; Li *et al*., 2022; Pei *et al*., 2022).

Ocorre que a GPP é influenciada por múltiplas variáveis (Li *et al*. (2022); e, assim sendo, muitos modelos de análise da GPP foram propostos (Pei *et al*., 2022; Zheng *et al*., 2020; Zhang *et al*., 2022; Li *et al*., 2022; Dias *et al*., 2024). Para esse último (Dias *et al*., 2024), há limitações nos modelos lineares tradicionais; segundo Zheng *et al*. (2020), os modelos propostos têm problemas principalmente na quantidade de dados; e, os muitos produtos globais desenvolvidos usando modelos Modelos de eficiência de uso de luz (LUE) têm cobertura temporal relativamente curta e não são atualizados. Suas limitações impedem de monitorar a dinâmica recente do carbono da vegetação global e fornecer bases científicas para a tomada de decisão.

Zheng *et al*. (2020) destacaram ainda que os modelos baseados em satélite têm sido amplamente utilizados para simular GPP de vegetação no local, regional ou global nos últimos anos. No entanto, tais simulações são ainda um grande desafio (Pei *et al*., 2022; Zheng *et al*., 2020). Desde os anos 2000, são produzidos dados de sensoriamento pelo *Moderate-resolution Imaging Spectroradiometer* (MODIS), um dos cinco sensores lançados pela *National Aeronautics and Space Administration* (NASA) (Anderson *et al.*, 2022). Suas contribuições têm sido de grande avanço para pesquisadores e para a humanidade.

O uso de SHAP com dados de GPP no Brasil é uma novidade acadêmica porque une ciência ambiental e inteligência artificial explicável (XAI), permitindo entender como e por que variáveis ambientais influenciam a produtividade das plantas. Essa abordagem ainda é rara no Brasil, especialmente em escala nacional e aplicada a biomas diversos. Além disso, o SHAP traz transparência aos modelos de *Machine Learning* (ML), contribuindo para decisões mais confiáveis e políticas públicas baseadas em evidências.

1.2. OBJETIVOS

1.2.1 Objetivo Geral

Aplicar modelos do tipo Comitê regressivo e *SHapley Additive exPlanations* (*SHAP)*, com fins de mensurar o impacto de fatores de influência, focos de incêndio e queimadas na GPP do cerrado brasileiro.

1.2.2 Objetivos Específicos

- Conhecer a GPP mensal do cerrado brasileiro da série histórica de 2003 a 2020;
- Conhecer os indicadores climáticos no cerrado brasileiro da série histórica de 2003 a 2020;
- Conhecer os focos de incêndio e sua abrangência na região do cerrado brasileiro da série histórica de 2003 a 2020;
- Estabelecer a relação entre cada variável e a GPP da série histórica de 2003 a 2020;
- Analisar como os modelos do Tipo Comitê performam em relação a GPP e as variáveis relacionadas ao Tempo, Clima e Vegetação e Focos de incêndio, explicadas pelo valor SHAP podendo ser alternativas viáveis aos

modelos lineares, especialmente em análise ambiental, contribuindo assim para identificar e quantificar os principais fatores associados.

Neste estudo foi realizada uma análise exploratória por meio de IA, ML, modelos do tipo Comitê regressivo e *SHapley Additive exPlanations* (*SHAP),* fundamentada em uma revisão de literatura sobre a influência das variáveis **(**Ano e Mês do evento, Temperatura, Precipitação, Pressão Atmosférica, Temperatura Umidade no ar, Velocidade de vento, Radiação solar, Índice e Densidade de vegetação, Luz absorvida e Área e Focos de incêndio e queimadas) sobre a GPP de uma amostra representativa do Cerrado brasileiro.

O SHAP quando aplicado apresenta algumas vantagens interpretabilidade, consistência com Teoria dos Jogos, contribuições marginais, comparabilidade, identificação de Interatividade e aplicabilidade em modelos complexos como neste estudo, sendo uma uma ferramenta valiosa para pesquisadores que desejam explorar a dinâmica subjacente e as influências no GPP, permitindo uma tomada de decisão mais informada em gestão ambiental, conservação e política ecológica (Lundberg e Lee, 2017).

2. REVISÃO DE LITERATURA

A Revisão de literatura neste estudo está dividida em três capítulos: o primeiro incluiu conceitos sobre a GPP, as variáveis de influência; o segundo situa o leitor sobre o campo de estudo - o Cerrado Brasileiro, sobre incêndios e Queimadas; e o terceiro capítulo sobre os Modelos Explicativos fundamenta o entendimento do leitor sobre os métodos e análise conceitual e aplicável ao contexto do estudo sobre IA e XAI, ML, SHAP, Métodos de Aprendizado Supervisionado. de Regressão, Modelos Baseados em Árvores, Árvores Aleatórias (*Random Forest*), Extremamente Aleatórias (*Extra Trees*), de Comitê (*Ensemble*), Adaptive Boosting (*AdaBoost*) e de Avaliação.

2.1. CAPÍTULO 1. *GROWTH PRIMARY PRODUCTION* (GPP) E OS FATORES DE INFLUÊNCIA

A Fotossíntese é um fenômeno natural imprescindível para a vida no planeta [ref]. É um mecanismo responsável pelo fluxo global de carbono e se associa a diversas atividades e acontecimentos no ecossistema como: respiração (R), crescimento vegetal, e principalmente a assimilação do carbono atmosférico (CO^2), e consequente produção de compostos orgânicos pelas plantas, também conhecida como GPP. Essa, juntamente com a R, compõem o principal processo que controla as trocas de CO^2 entre os ecossistemas terrestres e a atmosfera (Pei *et al*., 2022).

A GPP é definida como a quantidade de carbono fixada pelos ecossistemas terrestres por fotossíntese, e é considerada a base para produção[2] de vegetal. Uma fração substancial do carbono fixado pela GPP é perdida pela R do ecossistema. O restante é destinado à biomassa estrutural dos caules, folhas e frutas. Assim sendo a GPP tornou-se um importante parâmetro biofísico e pode auxiliar na compreensão e na dinâmica dos fluxos de carbono (Pei *et al*., 2022), bem como no ciclo vital representado na Figura 1.

[2] Produção é o processo pelo qual dois ou mais insumos são combinados para formar um novo produto. Por exemplo, nutrientes do solo, água, dióxido de carbono (CO^2) e luz solar são combinados para formar matéria orgânica durante a fotossíntese.

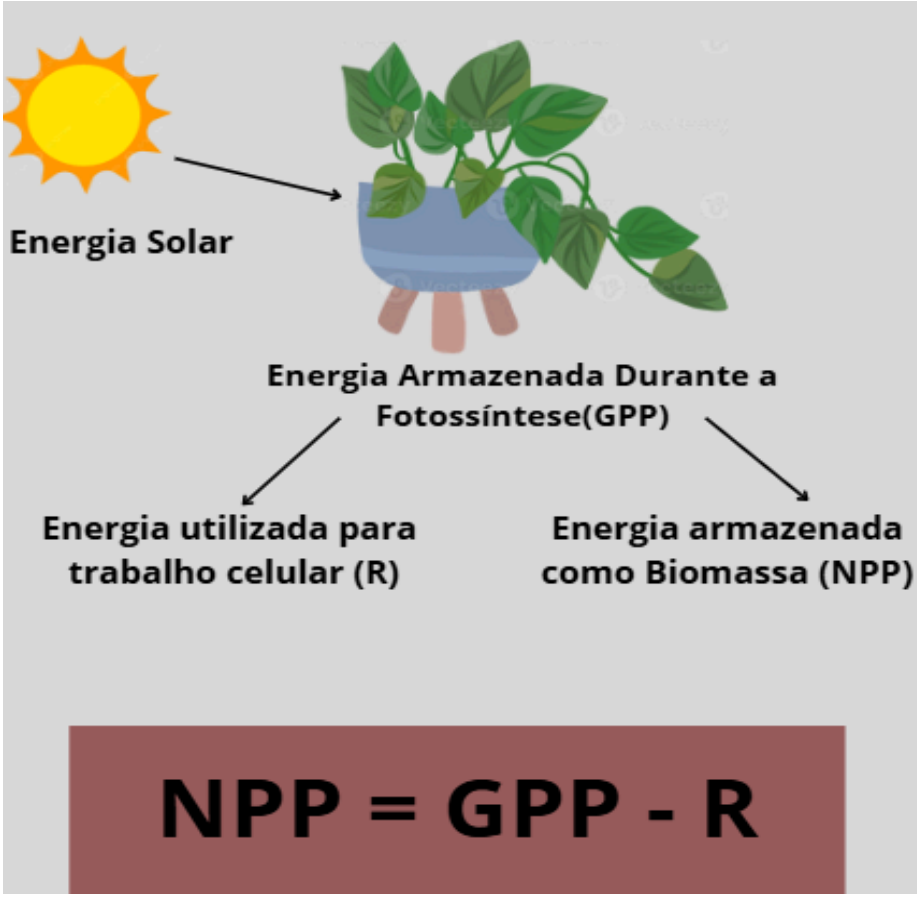

Figura 1. Fluxo de Energia
Fonte: Próprio Autor

A Figura 1 ilustra o processo de produtividade primária nos ecossistemas, que ocorre a partir da luz do sol, principal fonte de energia, que é captada pelas plantas na fotossíntese. Durante esse processo, a energia solar é convertida em energia química e armazenada em compostos orgânicos, sendo essa quantidade total denominada GPP. Parte dessa energia capturada é utilizada pelas plantas para seu próprio funcionamento e manutenção, como respiração celular, crescimento e reparo de tecidos, e é representada pela R. O restante da energia, após descontada a utilizada na R, é armazenada na forma de biomassa, disponível para os consumidores primários (herbívoros), sendo essa energia chamada de NPP, a energia disponível para o ecossistema resultante da subtração da energia gasta na R da GPP representada Equação 1.

Equação 1: $(GPP = NPP + R)$

A GPP é capaz de integrar fatores climáticos, ecológicos, geoquímicos e de influência humana sobre o planeta, e pode variar consideravelmente no espaço e no tempo. No espaço contribuem para as variações os fatores climáticos, a distribuição da vegetação e o uso do solo. No que se refere ao Tempo destaca-se como fatores de influência a respostas fisiológicas, as mudanças na atmosfera e a redistribuição dos ecossistemas. Mudanças e seus impactos e/ou de pressão sobre o meio natural, promovida pela ação antrópica, como a agricultura, em um determinado ambiente também são fatores de influência (De Santana, Delgado e Schiavetti, 2020).

Além disso, na atualidade dos estudos holísticos e integrados sobre a GPP emerge um importante papel em relação às questões econômicas, sociais e ambientais, e várias variáveis ambientais devem ser incluídas nos modelos de estudo da GPP. Em primeiro lugar destaca-se que o aumento da concentração de CO^2 na atmosfera nas últimos décadas estimularam substancialmente o crescimento da vegetação global; em segundo lugar, a radiação solar ou ativo fotossintético radiação (PAR) influencia substancialmente a produção vegetal do ecossistema terrestre; em terceiro lugar, o *Déficit* de pressão de vapor atmosférico (VPD) reflete o poder de secagem do ar sobre das plantas (Zheng *et al*., 2020).

Também há que se destacar sobre o ciclo de CO^2 terrestre a questão do aquecimento global, que intensifica as discussões sobre as mudanças climáticas e o real impacto do aumento das concentrações de GEE, em especial o CO^2, nos diversos ecossistemas terrestres, que pode resultar no desequilíbrio climático a nível global. Uma das principais fontes de emissão desses gases são atribuídas ao desmatamento, à agricultura e a outros usos do solo, que são responsáveis por cerca de 22% das emissões mundiais de GEE (De Santana, Delgado e Schiavetti, 2020).

Por isso, alerta-se que em cenários de aumento das emissões de CO^2, a capacidade de assimilação dos ecossistemas, embora aumente, a eficácia diminui. Nesse contexto, os estudos dos efeitos da mudança de uso e ocupação do solo nas trocas de energia, água e carbono entre a superfície e a atmosfera são fundamentais para conhecer a dinâmica desses processos e a capacidade dos ecossistemas terrestres de absorver o CO^2 excedente (De Santana, Delgado e Schiavetti, 2020).

Segundo Zhang *et al*. (2022), atualmente, os métodos de estimativa do GPP podem ser divididos principalmente em quatro tipos: o primeiro é o método empírico,

que estima o GPP de acordo com relações empíricas ajustadas usando observações limitadas e localizadas; o segundo método é baseado em LUE ou fluorescência de clorofila induzida por energia solar; no terceiro método consiste num modelo baseado em processos, como o modelo *Base Erosion Profit Shifting* (BEPS); e, por fim o ML, ou seja, modelos GPP orientado por dados, como floresta aleatória e aprendizado de transferência em várias etapas, um método que requer muitas observações do local, e a precisão da GPP estimada pode ser bastante reduzida se o conjunto de dados de treinamento não for grande o suficiente (Li *et al*., 2022). Em geral, os modelos GPP baseados em LUE orientados por sensoriamento remoto ainda são mais usados na estimativa de GPP em várias escalas devido à sua maior disponibilidade de dados e menos parâmetros.

Segundo Zheng *et al*. (2020), num estudo global de longo prazo de um conjunto de dados da GPP usando um Modelo de eficiência de uso de luz (EC-LUE), que integrou várias variáveis ambientais importantes a longo prazo, incluindo o modelo atmosférico CO^2 concentração, componentes de radiação, e pressão atmosférica de vapor d'água, demonstrou que os muitos produtos globais desenvolvidos usando modelos LUE têm cobertura temporal relativamente curta (normalmente menos de 20 anos) e não são atualizados no tempo. Essas limitações impedem de monitorar a dinâmica recente do carbono da vegetação global e fornecer bases científicas para a tomada de decisão. As variáveis ambientais mostraram mudanças substanciais de longo prazo e contribuíram significativamente para a vegetação em escala interanual. O modelo EC-LUE revisado teve um bom desempenho simulando as variações espaciais, sazonais e interanuais da GPP. Em particular, tem uma superioridade única na reprodução das variações interanuais. O modelo EC-LUE forneceu uma estimativa alternativa integrando as importantes variáveis, que a longo prazo podem ser bem refletidas no GPP global.

Li *et al*. (2022) aplicaram um modelo e uma técnica de interpretação de ML de alta fidelidade para desvendar os efeitos das variáveis climáticas na Produtividade primária líquida média (PPLM), na floresta amazónica. O modelo *Extreme Gradient Boosting* (XGBoost)[3] foi empregado para modelar os dados de

[3] Biblioteca de ML desenvolvido por Tianqi Chen da Universidade de Washington (USA), distribuída e de código aberto que utiliza árvores de decisão com reforço gradativo, um algoritmo de aprendizado supervisionado que faz uso do gradiente descendente. É conhecido por sua velocidade, eficiência e capacidade de escalar bem com grandes conjuntos de dados (Ludermir, 2021).

GPP do *Moderate-resolution Imaging Spectroradiometer* (MODIS), e o método de explicação aditiva SHAP foi introduzido para contabilizar as relações não lineares entre variáveis identificadas pelo modelo. O fator dominante da GPP na floresta amazônica variou em diferentes regiões, com a temperatura dominando a maior parte da ecorregião com um alto índice de importância. Além disso, o aumento da luz, o aumento da concentração de CO^2 e a diminuição da precipitação contribuíram positivamente para a GPP. A velocidade do vento para a maioria das áreas vegetadas estava abaixo do ótimo, o que beneficia a GPP, enquanto uma alta velocidade do vento sustentada traria uma perda substancial da GPP. Uma resposta não monótona da GPP ao VPD foi atribuída à carga de umidade. Como a GPP é influenciada por múltiplas variáveis climáticas, adotou-se uma técnica de ML explicável para entender como a GPP médio de longo prazo do MODIS respondeu às variáveis climáticas. A operação específica incluiu: primeiro, a GPP foi aprendida através de um potente modelo de ML; em seguida, o modelo foi alimentado no *framework* SHAP para explicar os mecanismos. As contribuições relativas de cada fator foram identificadas, mostrando que a temperatura superou outras variáveis climáticas na contribuição para a variabilidade da GPP; radiação e *déficit* de pressão de vapor também tiveram uma contribuição considerável; velocidade do vento, concentração de CO^2 e precipitação também. Na maioria das áreas, a temperatura excedeu o valor ótimo para o crescimento da GPP. Em geral, radiação elevada e aumento da concentração de CO^2 promovem ganhos de GPP, ao contrário da precipitação. Além disso, para a maioria da vegetação, a velocidade do vento não alcançou o valor ótimo que beneficia a GPP, e a alta velocidade do vento sustentada trouxe perda substancial de GPP. Considerando a resposta distinta da GPP ao conteúdo de água do solo em diferentes camadas, a relação entre a GPP e o VPD estave altamente conectada à política de uso da água e às condições de sobrecarga de umidade na floresta amazônica; e, os aumentos adicionais no VPD prejudicaram significativamente a GPP, apesar das condições de sobrecarga de umidade na floresta amazônica.

Pei *et al*. (2022) destacaram a importância dos fatores meteorológicos como temperatura e umidade na otimização do padrão espacial e temporal da GPP num estudo com LUE. Também foram identificados incertezas sobre o modelo e uso da luz, parametrizações e dados de entrada com muitas resoluções e precisões e

incompatibilidade com os dados do sensoriamento remoto e das torres de fluxo e observações de torres de fluxo. Os indicadores relacionados à refletância fotoquímica, a fluorescência da clorofila induzida pelo sol e a reflectância da vegetação no infravermelho, simplificam os métodos para estimar a GPP, mas não conseguem separar as influências de diferentes fatores ambientais. Espera-se que essas descobertas sobre a evolução dos modelos LUE e suas incertezas contribuam para futuras melhorias no modelo.

Para Dias *et al*. (2024), há limitações nos modelos lineares tradicionais em relação ao ML (RF e *XGBoost*) na captura de padrões intrincados, e essa oferece compreensão refinada da dinâmica das relações complexas e não lineares entre variáveis ambientais, desmatamento e fatores socioeconômicos e propõe estratégias mais eficazes para conservação e gestão sustentável da terra. Os autores conduziram uma análise exploratória abrangente de dados e apresentaram uma abordagem inovadora usando técnicas de ML interpretáveis e valores de *Shapley*, para mostrar como a IA explicável (XAI) pode servir como uma alternativa viável para modelos lineares, especialmente em análise ambiental da Amazônia, entre 1999 a 2020.

2.2. CAPÍTULO 2**.** O CERRADO BRASILEIRO E AS QUEIMADAS

O Cerrado brasileiro (Figura 2) se relaciona à uma grande extensão das savanas brasileiras, o segundo maior bioma dos seis existentes no Brasil, o qual possuía uma área territorial original de dois milhões de Km^2, ocupando 22% de todo território do país. Possui uma imensa abrangência, se estendendo pela região central do país e por 16 estados brasileiros: Região Norte, Tocantins (TO), Piauí (PI) e Rondônia (RO), Região Centro-Oeste Goiás (GO), Mato Grosso (MT), Mato Grosso do Sul (MS) e Distrito Federal (DF), da Região Nordeste, Bahia (BA), Ceará (CE) e Maranhão (MA) e da Região Sudeste, São Paulo (SP) e Minas Gerais (MG). Os estados da região norte como Roraima (RR), Amapá (AP), Amazonas (AM) e Pará (PA) também possuem a presença do Cerrado, porém na forma de manchas isoladas (Sawyer *et al*., 2017).

É o maior *hotspot*[4] do Hemisfério Ocidental, cobrindo mais de dois milhões de km^2 no Brasil e partes menores (cerca de 1%) da Bolívia e do Paraguai. Seu bioma é o segundo maior bioma da América do Sul, cobrindo uma área de 2.039.386 km2, 24% do território do Brasil (Sawyer *et al*., 2017).

[4] *Hotspot* (termo em inglês que significa lugar quente) foi criado, em 1988, pelo ecólogo inglês Norman Myers, para identificar áreas prioritárias para a biodiversidade. Concentram alta biodiversidade, associada a uma grande ocorrência de endemismos e sujeitas a grande pressão antrópica (ações do homem: desmatamento, as queimadas, a poluição, a caça e a pesca ilegais, o tráfico de animais, a destruição de habitats, a introdução de espécies exóticas, dentre outras.), à extinção de espécies animais e aos impactos das mudanças climáticas globais. Podem estar ameaçadas de extinção ou destruição e por essa razão são áreas indicadas pelos especialistas como prioritárias para serem protegidas e conservadas. Uma área é considerada hotspot quando tem pelo menos 1.500 espécies endêmicas de plantas e tenha perdido mais de ¾ (três quartos) de sua vegetação original.

Figura 2. Localização da área de estudo, Mapa do Bioma brasileiro
Fonte: https://www.ihuonline.unisinos.br/artigo/6749-biomas-brasileiros-e-a-teia-da-vida

Costa (2022) realizou uma análise bibliométrica sobre as publicações do Cerrado e o fogo dos últimos 20 anos usando dados da *Web of Science* e *VOSViewer*, *softwares* específicos para esse tipo de análise. Os documentos que mais se destacaram estão relacionados à distribuição do Cerrado e os principais *drivers* que controlam essa distribuição. O CO^2 e as consequências das mudanças climáticas também são protagonistas nos documentos mais influentes. Além disso, observou-se que os ecólogos ainda estão em busca de muitas respostas, e portanto é necessário a mobilização de ecólogos para se aprofundarem ainda mais nos estudos do fogo no Cerrado.

As queimadas podem ser utilizadas para conduzir modelos regionais de emissões e transporte de gases e química atmosférica. Alguns de seus importantes impactos incluem: a mudança do estado físico da vegetação, a liberação de gases de efeito estufa e de gases reativos durante a queima da biomassa; e outros particulados, ocasionando mudanças nas trocas de energia e água entre a superfície e a atmosfera, bem como alterações na comunidade vegetal devido a alterações no solo, como temperatura e mistura de componentes (Anderson *et al*., 2022).

2.3. CAPÍTULO 3. MODELOS EXPLICATIVOS

2.3.1 Inteligência Artificial (IA) e Inteligência Artificial Explicável (XAI)

Turing (1950), no artigo *"Computing Machinery and Intelligence"*, lançou as bases para o estudo da IA ao questionar se as máquinas poderiam pensar. Diante da dificuldade de definir o que se entendia por "pensamento", o autor propôs o Teste de Turing, um experimento no qual uma máquina seria considerada inteligente caso pudesse imitar o comportamento humano numa conversa textual. A inteligência deveria ser avaliada pelo desempenho observável, e não pela introspecção ou consciência, antecipando conceitos fundamentais como ML e redes neurais.

Russell e Norvig (2016), no livro “Inteligência Artificial: Uma Abordagem Moderna”, definiram a IA como sendo o estudo de agentes que recebem percepções do ambiente e realizam ações que maximizam o sucesso. Destacaram também que a IA envolve a simulação de processos cognitivos, como aprendizado, raciocínio e resolução de problemas em diversas disciplinas, incluindo ciência da computação, matemática e estatística. Dessa forma, a abordagem reforça a ideia de que a IA denota-se como campo multidisciplinar voltado para o desenvolvimento de sistemas capazes de reproduzir habilidades associadas à inteligência humana, tornando-se, assim, uma ferramenta para enfrentar desafios complexos e automatizar tarefas.

A IA se apresenta fortemente atrelada ao uso de modelos e algoritmos complexos, os quais exigem grande poder computacional de processamento, como redes neurais profundas e métodos *ensemble*, os quais frequentemente operam como sistemas *Black Box*. Esses modelos se mostram caracterizados por sua alta capacidade preditiva, porém, ao mesmo tempo, apresentam baixa interpretabilidade, dificultando a compreensão sobre como e quais se mostram as decisões tomadas por esses modelos. De acordo com Doshi-Vélez e Kim (2017), no artigo *"Towards a Rigorous Science of Interpretable Machine Learning"*, os sistemas *Black Box* oferecem precisão estatística impressionante, mas carecem de transparência e interpretabilidade, tornando-se um desafio em aplicações críticas, como diagnóstico médico e decisões judiciais.

Para lidar com essa nebulosidade dos modelos *Black Box*, surgem iniciativas como a XAI, a qual busca criar técnicas para explicar o funcionamento interno desses modelos Caixa Preta. Segundo o relatório da *Defense Advanced Research Projects Agency* (DARPA) (2016), a XAI visa aumentar a interpretabilidade sem comprometer o desempenho preditivo, tornando possível identificar vieses e erros ocultos nos modelos. Outrossim, estudos como o de Ribeiro, Singh e Guestrin (2016), *"Why Should I Trust You? Explaining the Predictions of Any Classifier"*, destacam o desenvolvimento de ferramentas como o *Local Interpretable Model-Agnostic Explanations* (LIME), que permite gerar explicações locais e compreensíveis para modelos complexos.

2.3.2 Valores Shapley

Os valores de Shapley foram formalmente definidos por Lloyd S. Shapley em seu artigo *"A Value for n-Person Games"*, publicado na obra *Contributions to the Theory of Games* (1952). Esses valores se mostram como uma solução para distribuir de forma justa os ganhos resultantes de uma coalizão em jogos cooperativos, baseando-se na contribuição marginal esperada de cada jogador. A fórmula matemática proposta calcula a importância individual considerando todas as permutações possíveis de participação, garantindo propriedades como eficiência, simetria, adicionalidade e consistência (Shapley, 1953).

O livro "*Theory of Games and Economic Behavio*r", publicado em 1944 por John Von Neumann e Oskar Morgenstern, se apresenta como um exemplar fundamental para o desenvolvimento da Teoria dos Jogos, estabelecendo os princípios matemáticos visando modelar interações estratégicas entre agentes racionais. Nele, introduziu-se conceitos de jogos cooperativos e não-cooperativos, destacando a diferença entre situações em que os jogadores podem formar coalizões para maximizar ganhos coletivos e aquelas em que agem individualmente para atingir seus objetivos (Sousa, 2005).

O jogo de coalizão se mostra como um modelo matemático da Teoria dos Jogos, no qual um jogadores pode formar grupos para cooperar e alcançar um determinado benefício coletivo. Esse modelo se concentra na alocação dos ganhos

ou custos resultantes da cooperação entre os jogadores, analisando como esses recursos devem ser distribuídos de forma justa entre os membros do grupo. A obra também mostra aspectos sobre a formalização de jogos de coalizão, designando as bases para a distribuição justa de ganhos entre os participantes, o que inspirou desenvolvimentos posteriores, como os valores de Shapley (Sousa, 2005).

2.3.3 *SHapley Additive exPlanations* (SHAP)

No artigo "*A Unified Approach to Interpreting Model Predictions*" de Lundberg e Lee (2017), o SHAP se apresenta definido como um método unificado para interpretar modelos preditivos. Ele se baseia na teoria dos jogos, utilizando os valores de Shapley para atribuir uma importância a cada característica (*feature*) na previsão de um modelo.

O *SHAP* se apresenta descrito como uma abordagem aditiva de atribuição de importância das características (*feature importance*) que atende a três propriedades fundamentais:

1. Precisão Local (*Local Accuracy*): O modelo explicativo visa corresponder à saída do modelo original para uma entrada específica;
2. Ausência (*Missingness*): Características ausentes não devem ter impacto nas previsões.
3. Consistência (*Consistency*): Se uma determinada característica aumenta sua contribuição em um modelo, seu valor de importância nunca deve diminuir.

Sobre essa ótica, a explicação do valor de Shapley se mostra como um método aditivo de atribuição de recursos, um modelo linear. Um modelo de explicação, sendo função linear de variáveis binária (EQUAÇÃO 2):

Equação 2: $$g(z') = \Phi_0 + \sum_{i=1}^{M} \Phi_i z_i'$$

- $g(z')$: Função de explicação do modelo, que estima a predição com base nas características explicativas.

- ϕ_0: Valor esperado da predição quando todas as características estão ausentes, atuando como uma constante ou ponto de base.
- ϕ_i: Valores atribuídos (importância) a cada característica i. Representa a contribuição marginal de cada variável no modelo.
- $z_i^{'}$: Variável binária (0 ou 1) que indica a presença ou ausência de uma característica no conjunto analisado.
- M: Número total de características analisadas no modelo.

Lundberg e Lee (2017) definiram que o SHAP se mostra como um método essencial dentro da abordagem de XAI, uma vez que permite que modelos *Black Box* sejam interpretáveis ao calcular e justificar a importância de cada entrada para a previsão. Dessa forma, a abordagem condicionada pelos autores catalisa uma maior transparência e ajuda a identificar vieses ou erros nos modelos.

O SHAP é uma técnica avançada de interpretação de modelos que pode ser utilizada para entender a contribuição individual de cada característica (*feature*) em modelos preditivos, especialmente os ML. Utilizar o SHAP para medir o GPP (em modelos ambientais ou ecológicos pode trazer várias justificativas válidas como:

- Interpretabilidade: O SHAP fornece uma forma clara e consistente de interpretar os efeitos de cada característica nos modelos de previsão de GPP. Isso pode ajudar a entender melhor quais fatores são mais influentes na produtividade primária bruta de um ecossistema.
- Consistência com Teoria dos Jogos: O SHAP baseia-se na teoria dos valores de Shapley da teoria dos jogos, que atribui um valor justo de importância a cada característica com base em sua contribuição para todas as possíveis combinações de características.
- Contribuições Marginais: Considera as contribuições marginais de cada característica, permitindo uma análise mais granular de como diferentes variáveis ambientais (como temperatura, umidade, luz solar, etc.) afetam o GPP.
- Comparabilidade: Os valores SHAP podem ser comparados diretamente entre os modelos, o que facilita a análise de diferentes modelos de previsão de GPP ou a comparação de diferentes regiões ou condições ambientais.

- Identificação de Interatividade: Também pode identificar interações entre características que não são facilmente visíveis em outros métodos de interpretação de modelo. Essa capacidade de detectar interações complexas é particularmente útil em estudos ecológicos onde fatores como interações entre espécies ou *feedbacks* ambientais são significativos.
- Aplicabilidade em Modelos Complexos: Modelos que preveem GPP frequentemente utilizam técnicas de ML complexas, como florestas aleatórias ou redes neurais, as quais são possíveis de serem decompostas, tornando-os mais acessíveis para compreensão humana (Lundberg e Lee, 2017).

2.3.4 Estimativa Clássica do Valor de Shapley

A estimativa clássica do valor de Shapley, conforme descrita por Lipovetsky e Conklin, em 2001, se apresenta como uma abordagem utilizada para determinar a importância relativa de cada variável de entrada em modelos de regressão. Esse método envolve o reprocessamento do modelo em todos os subconjuntos possíveis de variáveis $S \subseteq F$, onde F representa o conjunto completo de variáveis disponíveis. O objetivo é avaliar o impacto marginal que a inclusão ou exclusão de cada variável tem na previsão do modelo. Para isso, o modelo é treinado tanto com a variável presente $f(S \cup \{i\}$ quanto com a variável ausente ($f(S)$), e as previsões resultantes são comparadas. O valor de Shapley ($\Phi_i(f, x)$) é calculado como a média ponderada dessas diferenças ao longo de todos os subconjuntos analisados, de acordo com a fórmula (Lipovetsky e Conklin, 2010) (EQUAÇÃO 3).

Equação 3: $$\Phi_i(f, x) = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F|-|S|-1)!}{|F|!} [f(S \cup \{i\}) - f(S)]$$

- $\Phi_i(f, x)$: Valor de Shapley atribuído à característica i. Representa a contribuição marginal esperada dessa característica para o resultado do modelo.
- F: Conjunto de todas as características disponíveis.
- $f(S)$: Valor da função modelo (predição) usando apenas o subconjunto S.

- $|S|!$: Fatorial do tamanho do subconjunto S
- $S \subseteq F \setminus \{i\}$: Subconjuntos possíveis das características sem incluir a característica i.
- $|S|!\,(|F| - |S| - 1)!$: Fatorial do número de características restantes após incluir S e a característica i.
- $|F|!$: Fatorial do número total de características, normalizando o cálculo
- $f(S \cup \{i\})$: Valor da função modelo ao incluir a característica i no subconjunto S.

Esse método adota uma abordagem aditiva, distribuindo as contribuições das variáveis com base em sua influência sobre o modelo. Outrossim, permite o uso de representações binárias (0 ou 1), visando indicar a inclusão ou exclusão de uma variável no modelo, tornando o processo mais intuitivo e didático (Lipovetsky e Conklin, 2010).

2.3.5 *Machine Learning* (ML)

Segundo Mitchell (1997), em seu livro *"Machine Learning"*, o ML é descrito como: um programa é dito aprender a partir da experiência E em relação a uma tarefa T e uma medida de desempenho P, se seu desempenho em T, medido por P, melhora com a experiência E. Essa definição destaca que o aprendizado ocorre quando um sistema utiliza dados históricos ou exemplos para ajustar automaticamente seus parâmetros e generalizar padrões, permitindo previsões ou decisões mais precisas em novos cenários.

Alpaydin (2020), em seu livro *"Introduction to Machine Learning"*, complementa que o ML utiliza métodos estatísticos e matemáticos para identificar padrões em grandes volumes de dados e, assim, automatizar tarefas como classificação, regressão, agrupamento e reconhecimento de padrões.

O principal objetivo do ML é resolver uma determinada tarefa específica para a qual foi previamente treinado (Flach, 2012). Dispondo-se a extrair informações relevantes de um domínio ou conjunto de dados, o processo inicial envolve a

identificação de atributos e variáveis que descrevem o problema em questão. Com base nesses atributos, define-se a tarefa a ser executada.

No contexto do ML, algumas das tarefas mais comuns incluem:

- Classificação: organizar os dados em categorias predefinidas, na qual cada classe representa um conjunto discreto de valores prévios;
- Regressão: Determinar uma função matemática que relacione os dados de entrada a um valor contínuo, possibilitando previsões;
- Agrupamento: Dividir os dados em grupos sem classes definidas previamente. Nesse caso, o objetivo é descobrir padrões naturais ou estruturas subjacentes nos dados.

A realização dessas tarefas depende do uso de um algoritmo de ML, que é treinado com um conjunto de dados representativos do domínio em questão. Esse conjunto contém os atributos identificados e serve como base para o algoritmo aprender padrões e construir um modelo preditivo. O modelo gerado é, em vista disso, o resultado da tentativa do algoritmo de identificar estruturas e relações nos dados de entrada. Caso o modelo utilize apenas um atributo como referência, ele é classificado como univariado. Se considera múltiplos atributos, é denominado multivariado (Flach, 2012).

2.3.6 Aprendizado Supervisionado e Não supervisionado

Os modelos de ML podem ser classificados com base na natureza do seu aprendizado, destacando-se, principalmente, os seguintes tipos: Aprendizado Supervisionado e Aprendizado Não-Supervisionado. No Supervisionado, o algoritmo se apresenta treinado utilizando pares de entrada e saída desejadas, visando o objetivo de aprender uma função de mapeamento que associa novas entradas aos valores corretos de saída. Entre as tarefas mais comuns desse tipo de aprendizado estão a Classificação e a Regressão (Flach, 2012).

Por outro lado, no Aprendizado Não-Supervisionado, o algoritmo não recebe rótulos ou saídas desejadas. Em vez disso, ele é responsável por identificar padrões e estruturas ocultas nos dados de entrada, agrupando elementos semelhantes com

base em características compartilhadas. Um exemplo clássico dessa abordagem é a tarefa de Agrupamento (Ludermir, 2021).

Os métodos de aprendizado Não-supervisionado podem ajudar a descobrir novos agrupamentos de dados, permitindo novas categorizações durante a rotulagem. A rotulagem de dados, ou anotação de dados, faz parte da etapa de pré-processamento no desenvolvimento de um modelo de ML.

Dados rotulados são usados em aprendizado Supervisionado, enquanto os dados não rotulados, os que praticamente não receberam anotação humana são usados em aprendizado Não-supervisionado. A rotulagem de dados requer a identificação de dados brutos (ou seja, imagens, arquivos de texto, vídeos) e, em seguida, a adição de um ou mais rótulos a esses dados para especificar seu contexto para os modelos, permitindo que o modelo de aprendizado de máquina faça previsões precisas. Os dados rotulados são mais difíceis de adquirir e armazenar (ou seja, consomem mais tempo e são caros), enquanto os dados não rotulados são mais fáceis de adquirir e armazenar. Os dados rotulados podem ser usados para determinar insights acionáveis (por exemplo, tarefas de previsão), enquanto os dados não rotulados são mais limitados em sua utilidade (Ludermir, 2021).

Destaca-se também outros tipos de aprendizado, como o Aprendizado Semi Supervisionado, o qual combina dados rotulados e não rotulados para melhorar o desempenho do modelo, e o aprendizado por reforço é um tipo de aprendizado de máquina que, diferentemente do aprendizado supervisionado e não supervisionado, não utiliza diretamente dados rotulados como entrada. Em vez disso, o algoritmo aprende por meio de tentativa e erro, interagindo com um ambiente e ajustando suas ações com base em recompensas ou penalidades recebidas.

Os modelos também podem ser categorizados de acordo com seus objetivos. Quando o foco se apresenta na previsão de valores a partir das variáveis de entrada, o modelo é denominado preditivo e geralmente resulta de algoritmos de aprendizado de máquina supervisionado. Por outro lado, quando o objetivo é a análise e descrição dos dados, sem a intenção de fazer previsões explícitas, o modelo é classificado como descritivo, sendo mais comum em algoritmos de aprendizado de máquina não supervisionado (Flach, 2012).

2.3.6.1 Métodos de Regressão

Montgomery, Peck e Vining (2021), no livro *"Introduction to Linear Regression Analysis"*, apresentaram os métodos de regressão como ferramentas estatísticas para modelar e quantificar a relação entre variáveis dependentes e independentes. Esses métodos se mostram aplicados não apenas para prever valores futuros, mas também para explicar como as variáveis preditoras influenciam a variável resposta, fornecendo *insights* sobre a importância relativa de cada fator. O foco da regressão está na construção de modelos matemáticos robustos, capazes de capturar padrões nos dados, minimizar os resíduos e garantir eficiência preditiva.

As variáveis dependentes são aquelas que representam o resultado ou resposta que se deseja prever, explicar ou analisar. Elas dependem diretamente das variáveis independentes, sendo influenciadas ou determinadas por essas.

- Definição segundo Montgomery, Peck e Vining (2021): A variável dependente é aquela que responde ou é influenciada por mudanças nas variáveis independentes e é utilizada como saída do modelo;
- As variáveis independentes, por outro lado, são os fatores ou preditores que influenciam diretamente a variável dependente. Elas são manipuladas ou observadas para verificar como afetam o resultado (Seber e Lee, 2003);
- As variáveis independentes, também conhecidas como variáveis explanatórias ou preditoras, são usadas para prever ou explicar as mudanças observadas na variável dependente (Seber e Lee, 2003).

Francis Galton, um polímata britânico do século XIX, introduziu o conceito de regressão em seus estudos sobre hereditariedade em relação a alturas de pais e filhos. Em seu trabalho "*Regression Towards Mediocrity in Hereditary Stature*" (1886), publicado no *Journal of the Anthropological Institute of Great Britain and Ireland*, Galton observou que, embora pais altos tendessem a ter filhos altos, as alturas dos filhos geralmente se aproximavam mais da média populacional, um fenômeno que ele denominou regressão em direção à mediocridade. Essa observação levou ao desenvolvimento do conceito estatístico de regressão à média, fundamental para a formulação da regressão linear (Galton, 1986).

A regressão linear é uma técnica estatística usada para modelar a relação entre uma variável dependente (Y) e uma ou mais variáveis independentes (X), assumindo que essa relação pode ser descrita por uma função linear : $Y = \beta 0 + \beta 1X + \epsilon$, onde $\beta 0$ é o intercepto, $\beta 1$ é o coeficiente angular (pendente) que indica a taxa de variação de Y em relação a X, e ϵ representa o termo de erro (EQUAÇÃO 4). O objetivo principal é prever valores ou entender a influência das variáveis independentes sobre a variável dependente (Figura 3).

Equação 4: $Y = \beta 0 + \beta 1X + \epsilon$

Y: Variável dependente (resultado que queremos prever).

X: Variável independente (fator preditor).

$\beta 0$: Intercepto (valor de Y quando X = 0).

$\beta 1$: Coeficiente angular (taxa de variação de Y em relação a X).

ϵ: Termo de erro aleatório, que captura variações não explicadas pelo modelo.

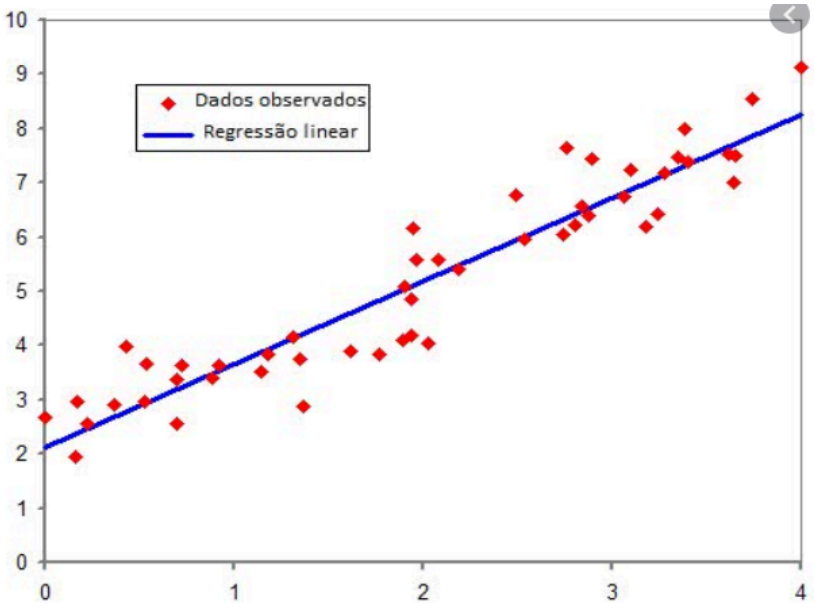

Figura 3. Gráfico Exemplo de Regressão Linear

A Figura 3 é um exemplo de aplicação da regressão linear, na qual os pontos vermelhos representam os dados, ou seja, os valores reais coletados, e a linha azul corresponde à reta de regressão, que melhor ajusta esses pontos com base no modelo de regressão linear. Seu objetivo é estabelecer uma relação linear entre uma variável independente (eixo X) é uma variável dependente (eixo Y), utilizando a equação $Y = \beta 0 + \beta 1X + \epsilon$. A linha azul é determinada de forma a minimizar os erros, que são as diferenças verticais entre os valores previstos pelo modelo (linha azul) e os valores observados (pontos vermelhos). A inclinação da

linha ($\beta 1$) indica a taxa de variação de Y em relação a X, enquanto o alinhamento geral dos pontos ao longo da linha reflete o grau de ajuste do modelo. Quanto mais próximos os pontos estiverem da linha, melhor o modelo se ajusta aos dados.

Apesar de ser um dos métodos de regressão mais antigos, introduzidos por Galton (1886), a regressão linear simples é ainda hoje um dos métodos amplamente utilizado em diferentes áreas do conhecimento, também sendo um dos métodos de regressão testados em nosso projeto.

2.3.6.2 Modelos Baseados em Árvores

Segundo Breiman *et al*. (1984), no livro *"Classification and Regression Trees (CART)"*, as árvores de decisão são estruturas hierárquicas criadas para modelar decisões com base em regras lógicas e testes binários. Nessas estruturas, cada nó interno representa uma pergunta sobre um atributo, enquanto cada ramificação corresponde a uma resposta para essa pergunta. Os nós terminais contêm as previsões finais ou classificações geradas pelo modelo. O funcionamento das árvores inicia-se na raiz, que é o primeiro ponto de decisão baseado em uma variável independente. A partir desse ponto, os dados são divididos em subgrupos menores nos nós internos, até chegarem às folhas, onde as previsões finais são realizadas. As divisões seguem critérios específicos, como : o Ganho de Informação (ID3), que avalia a redução da incerteza após a separação dos dados; o Índice de Gini (CART), que mede a pureza dos grupos formados; o Qui-quadrado (CHAID), que verifica a associação estatística entre os atributos; e a Razão de Ganho (C4.5), que aprimora o ganho de informação ao normalizar sua influência, reduzindo o viés em atributos com muitos valores distintos. As árvores de decisão podem ser aplicadas em dois principais contextos: Árvores de Classificação, utilizadas para prever categorias discretas, como exemplo clássico temos a análise de inadimplência de clientes no setor bancário, analisando as características daquele cliente se ele vai ou não pagar um empréstimo, como idade, salário mensal, quantidade de filhos, etc.; e, Árvores de Regressão, aplicadas para prever valores numéricos contínuos, como a estimativa do preço de um imóvel com base em suas características. Outro exemplo clássico na análise de variáveis para compreender se

um imovel se mostra mais caro pela quantidade de banheiros, tamanho em m^2, localização, etc. (Breiman *et al.*, 2001).

Os modelos de regressão baseados em árvores se apresentam como algoritmos preditivos que utilizam estruturas hierárquicas de árvores de decisão para modelar a relação entre variáveis independentes (preditoras) e uma variável dependente (resposta). Segundo Breiman *et al.* (2001), esses modelos dividem os dados em subconjuntos homogêneos por meio de divisões recursivas, minimizando o Erro quadrático médio (MSE) em cada etapa. Esse processo permite a construção de modelos flexíveis, capazes de lidar com relações não lineares e interações complexas entre variáveis (Figura 4).

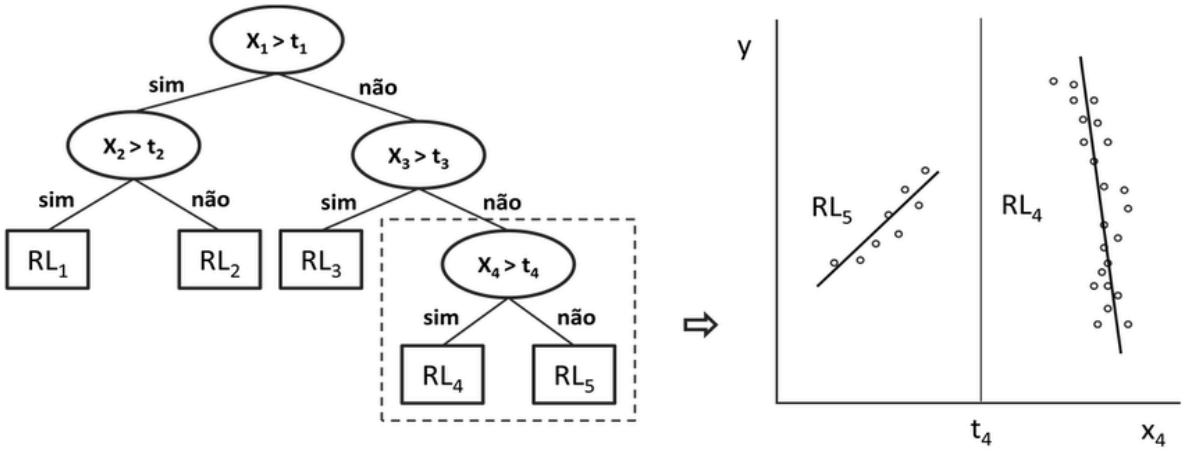

Figura 4. Árvores de regressão
Fonte:
https://www.researchgate.net/figure/a-Exemplo-de-Arvore-de-Regressao-com-quatro-atributos

A Figura 4 apresenta um exemplo de árvores de regressão, uma abordagem de aprendizado de máquina que é eficaz para capturar relações não lineares e interações complexas entre variáveis. No lado esquerdo da figura, é ilustrada a estrutura de uma árvore de decisão para regressão, composta por nós de decisão (representados por elipses) e nós terminais (representados por quadrados rotulados como RL1, RL2, etc.). Cada nó de decisão avalia uma condição sobre uma variável ($x_i > t_i$) e divide os dados em subgrupos com base em um valor limite (t_i). Por exemplo, o primeiro nó verifica se $x_i > t_i$, direcionando os dados para os ramos "sim" ou "não". Esse processo se repete até que os dados cheguem a um nó terminal, que representa o valor previsto para aquele subconjunto de dados.

No lado direito da figura, o gráfico demonstra como o espaço de atributos é dividido em regiões com base nas decisões tomadas pela árvore. Cada região (RL4, RL5) possui sua própria regressão linear ou valor constante previsto, ajustado de acordo com os dados

que caem nessa região. Essas divisões permitem que o modelo trate de maneira específica os diferentes padrões presentes no conjunto de dados.

2.3.6.3 Florestas Aleatórias (*Random Forest*)

As Florestas Aleatórias ou *Random Forest* (RF) são algoritmos de aprendizado supervisionado amplamente aplicados tanto em problemas de classificação quanto de regressão (Tibco, 2021). Esse método consiste na combinação de várias árvores de decisão, formando um modelo robusto e eficiente, capaz de reduzir problemas como o overfitting e aumentar a precisão preditiva (Breiman, 2001).

Conforme descrito por Breiman (2001), as árvores de decisão se apresentam como estruturas hierárquicas que funcionam como fluxogramas. Cada nó interno representa uma pergunta sobre um atributo dos dados, enquanto os ramos indicam as possíveis respostas e os nós terminais contêm as previsões finais ou classificações. A construção dessas árvores baseia-se em métricas como ganho de informação, entropia e CART para determinar os pontos de divisão mais eficazes nos dados. Contudo, o uso de apenas uma árvore pode resultar em sensibilidade a ruídos e *outliers*, comprometendo a capacidade do modelo de generalizar para novos dados (Figura 5).

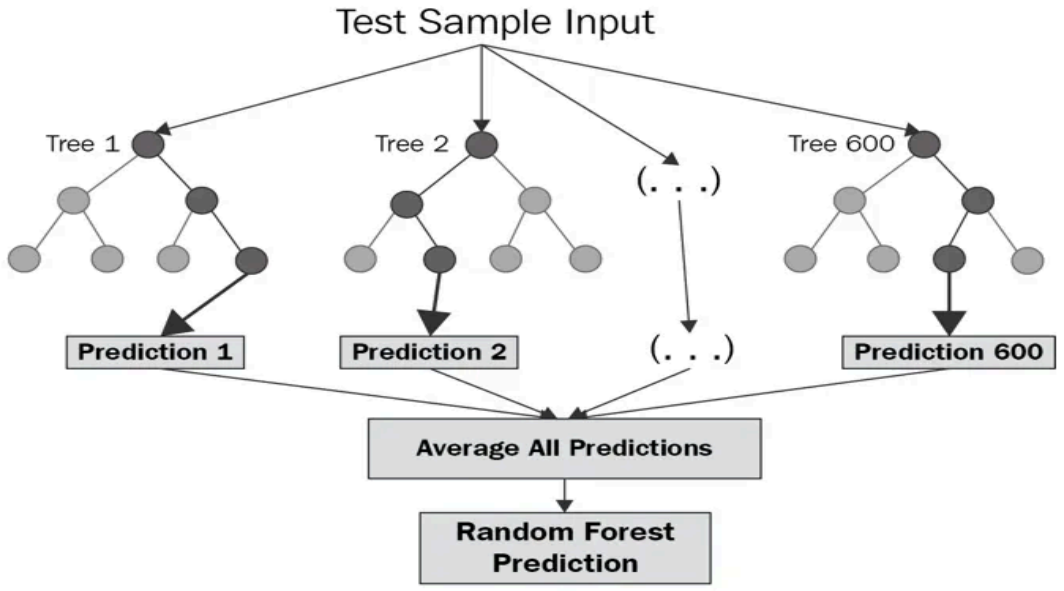

Figura 5. Exemplo de *Random Forest*
Fonte: https://corporatefinanceinstitute.com/resources/data-science/random-forest/

A Figura 5 ilustra o funcionamento do algoritmo RF, uma técnica de *ensemble learning* que utiliza múltiplas árvores de decisão para melhorar a precisão e reduzir a sensibilidade a ruídos. O processo começa com a entrada de um exemplo de teste, chamado *Test Sample Input*, que é avaliado por várias árvores de decisão independentes dentro do modelo. Cada árvore é construída com uma amostra aleatória do conjunto de treinamento, por meio da técnica de *bootstrap*, e faz sua predição de forma individual. No caso de problemas de regressão, as predições de todas as árvores são combinadas pela média, enquanto em problemas de classificação, o resultado final é determinado pelo voto majoritário entre as árvores. Ao combinar as predições de diversas árvores, o RF produz uma predição final mais robusta e confiável, reduzindo o impacto de ruídos e *outliers*. Essa abordagem melhora a generalização do modelo devido à aleatoriedade introduzida na seleção dos dados e dos atributos avaliados em cada nó das árvores, mitigando o risco de sobreajuste comum em modelos baseados em uma única árvore de decisão.

Para mitigar essas limitações, o RF utiliza um conjunto de árvores criadas com amostras aleatórias de dados e atributos (Flach, 2012). Esse método, conhecido como *ensemble learning*, combina os resultados das árvores individuais para formar uma previsão final. No caso de classificação, o resultado se mostra obtido por votação majoritária entre as árvores, enquanto para regressão, calcula-se a média das previsões.

Essa estratégia diminui a correlação entre as árvores e torna o modelo mais estável e preciso, reduzindo o risco de superajuste. O superajuste ocorre quando um modelo de aprendizado de máquina aprende detalhes e ruídos específicos do conjunto de treinamento, ao ponto de comprometer sua capacidade de generalizar para novos dados. Isso significa que o modelo tem um desempenho muito bom no treinamento, mas apresenta erros significativos em dados novos ou de teste. As RF também apresentam alta resistência a ruídos e *outliers*, o que as torna mais robustas do que árvores de decisão individuais. Isso é alcançado por meio da seleção aleatória de atributos em cada divisão e da amostragem *bootstrap*, garantindo maior diversidade entre as árvores e melhorando sua capacidade de generalização (Britto e Pacífico, 2020).

2.3.6.4 Árvores Extremamente Aleatórias (*Extra Trees*)

As Árvores Extremamente Aleatórias foram introduzidas por Geurts, Ernst e Wehenkel (2006), no artigo *"Extremely Randomized Trees"*, na revista *Machine Learning,* propondo um novo método de *ensemble* baseado em árvores para problemas de classificação e regressão supervisionados, e que consiste em randomizar fortemente tanto a escolha dos atributos quanto dos pontos de corte ao dividir o nó da árvore. O método constroi árvores totalmente aleatórias, cujas estruturas são independentes dos valores de saída do conjunto de treinamento. A força da randomização pode ser ajustada conforme as especificidades do problema por meio de parâmetros. Além da precisão, uma das principais vantagens do algoritmo resultante é a eficiência computacional.

O algoritmo ET é uma extensão do RF que se diferencia pelo maior grau de aleatoriedade introduzido na construção das árvores. Enquanto o RF utiliza amostragem *bootstrap* e escolhe o melhor ponto de corte com base em critérios de divisão como o ganho de informação, que depende de métricas como entropia ou razão de ganho, o ET vai além ao introduzir aleatoriedade adicional tanto na escolha das variáveis (atributos) quanto na escolha dos pontos de corte. Isso significa que, ao invés de buscar o ponto de corte ótimo, o ET escolhe os pontos de corte de forma totalmente aleatória dentro dos limites do atributo selecionado (Figura 6).

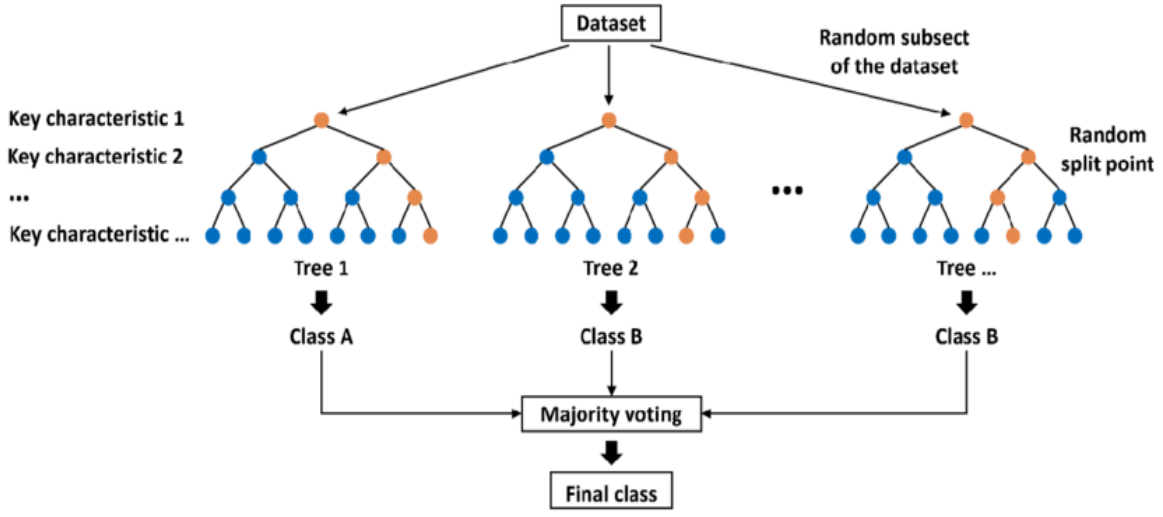

Figura 6. Extra Trees
Fonte: https://www.researchgate.net/figure/Structure-of-Extra-Trees-Kapoor-2020-Extra-Trees-constructs-the-set-of-decision-trees

A Figura 6 ilustra o funcionamento do algoritmo ET, introduzindo maior aleatoriedade na construção das árvores de decisão. O ET utiliza o conjunto completo de dados em vez de amostras e seleciona aleatoriamente um subconjunto de variáveis em cada nó de decisão. Além disso, os pontos de corte para essas variáveis são escolhidos de forma totalmente aleatória dentro dos limites possíveis, diferentemente do RF, que busca otimizar os pontos de corte com base em métricas como ganho de informação ou índice Gini. Cada árvore de decisão construída no ET gera sua predição individual, como Classe A ou Classe B. Para combinar as predições das árvores, o algoritmo utiliza a votação majoritária em problemas de classificação ou a média em problemas de regressão, resultando na classe ou valor final. Essa abordagem aumenta a diversidade das árvores, reduzindo a variância do modelo, e elimina a necessidade de cálculos de otimização para os pontos de corte, o que pode melhorar a eficiência computacional.

No ET, como os pontos de corte se mostram escolhidos aleatoriamente, isso tende a aumentar a variabilidade das árvores individuais e ajuda a reduzir a dependência do modelo em detalhes específicos dos dados de treinamento, resultando em melhor generalização para novos dados. Sobre essa ótica, torna-se possível uma maior robustez em dados ruidosos ou correlacionados, uma vez que a aleatoriedade dos pontos de corte no ET evita que o modelo depende excessivamente de atributos altamente correlacionados, tornando-o mais robusto para lidar com ruído e colinearidade nas variáveis preditoras (Geurts, Ernst e Wehenkel, 2006).

2.3.6.5 Modelos do Tipo Comitê

Os modelos do tipo Comitê ou também chamados de métodos *ensemble*, se apresentam como abordagens de ML que combinam múltiplos modelos para melhorar a precisão, reduzir o viés e a variância, além de aumentar a robustez e a estabilidade do modelo. Conforme Jadama e Toray (2024), essas técnicas incluem estratégias como *bagging*, que reduz a variância por meio de amostragem aleatória e combina previsões; *boosting*, que minimiza o viés ao corrigir erros

sequencialmente; e, *stacking*, que integra previsões de algoritmos heterogêneos usando um meta-modelo.

2.3.6.6 *Adaptive Boosting (Adaboost)*

O *AdaBoost* (Figura 7), desenvolvido por Yoav Freund e Robert Schapire (1995), denota-se um algoritmo de ML baseado na técnica de *boosting*, que combina múltiplos classificadores fracos para formar um classificador forte e mais preciso. Sua principal característica é a capacidade de ajustar iterativamente os pesos atribuídos às observações no conjunto de dados, dando maior ênfase às instâncias que foram classificadas incorretamente em etapas anteriores. Esse processo adaptativo permite que o algoritmo corrija erros, tornando-o altamente eficaz para tarefas de classificação. *AdaBoost* utiliza modelos simples, como árvores de decisão rasas (*stumps*), para construir cada classificador fraco e, ao final, combina os resultados por meio de um sistema de ponderação. Embora seja poderoso e flexível, o algoritmo é sensível a ruídos e *outliers*, o que pode afetar seu desempenho em conjuntos de dados desbalanceados. Sua eficácia e simplicidade o tornaram amplamente utilizado em aplicações práticas e pesquisas de aprendizado supervisionado (Freund e Schapire, 1995).

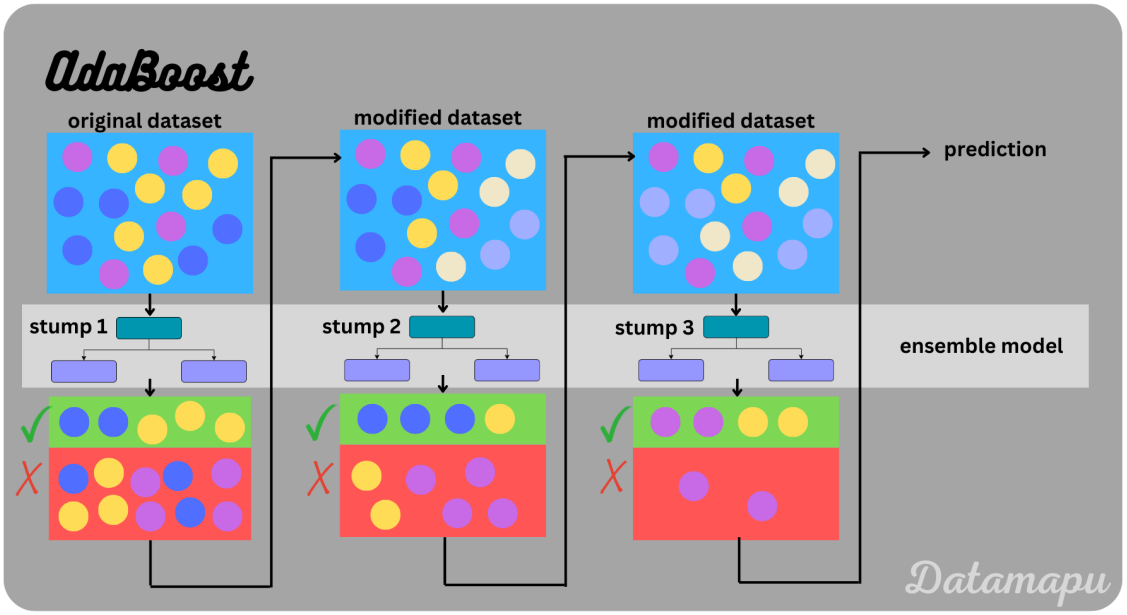

Figura 7. Exemplo do funcionamento do AdaBoost
Fonte: https://datamapu.com/posts/classical_ml/adaboost/

A Figura 7 ilustra o funcionamento do algoritmo AdaBoost, que é uma técnica de *ensemble learning* baseada em *boosting*. Inicialmente, o algoritmo utiliza um conjunto de dados original, onde todos os exemplos possuem o mesmo peso. Um modelo fraco, frequentemente uma árvore de decisão simples (ou *stump*), é treinado para classificar os dados. Após a primeira tentativa, os exemplos que foram classificados incorretamente recebem maior peso, destacando sua importância no próximo passo. Em seguida, o conjunto de dados é modificado, refletindo os pesos ajustados, e um segundo modelo fraco é treinado. Este processo é repetido iterativamente, com cada novo modelo se concentrando nos erros do anterior, ajustando os pesos dos exemplos incorretamente classificados a cada rodada. No final, os diferentes modelos fracos são combinados em um modelo de *ensemble*, onde cada modelo contribui para a predição final com um peso proporcional à sua precisão, resultando em um modelo forte e robusto que melhora significativamente a capacidade de classificação ou predição.

No artigo "*Improving Regressors Using Boosting Techniques*", Harris Drucker (1997) explora a aplicação de técnicas de *boosting* para aprimorar modelos de regressão. O estudo utiliza árvores de regressão como blocos fundamentais na construção de comitês de regressão, tanto para *bagging* quanto para *boosting* (Drucker, 1997).

2.3.7 Métodos de Avaliação

A avaliação do desempenho de modelos de aprendizado de máquina é essencial para garantir sua capacidade de realizar previsões precisas em novos dados. No caso específico de tarefas de regressão, onde o objetivo é prever valores contínuos, diversas métricas de erro são utilizadas para quantificar a qualidade das predições. O Erro Quadrático Médio (RMSE), equação 2.1, é destacado por sua capacidade de penalizar erros maiores, tornando-se útil em cenários onde desvios significativos precisam ser minimizados. Já o Erro Absoluto Médio (MAE) é enfatizado por sua simplicidade e robustez, fornecendo uma medida direta da magnitude dos erros, sendo menos sensível a *outliers*. O Erro Percentual Absoluto Médio (MAPE) é apresentado como uma métrica que expressa os erros em termos

percentuais, facilitando a interpretação relativa e permitindo comparações entre diferentes escalas de dados. Por outro lado, o MSE é analisado por ser semelhante ao RMSE, mas sem aplicar a raiz quadrada, ampliando a penalização de grandes erros e destacando padrões extremos nos dados. Por fim, o Coeficiente de Determinação (R^2) se apresenta descrito como uma métrica que avalia o grau de ajuste do modelo, indicando a proporção da variância explicada pelas variáveis independentes. A combinação dessas métricas fornece uma análise abrangente, permitindo identificar desempenhos robustos e fragilidades específicas em modelos preditivos, tornando-as indispensáveis na avaliação e otimização de algoritmos de ML. Para todas as equações o y_i: o valor real observado; $\hat{y}_i$: valor previsto observado e n: número total de observações, $\underline{y}$ se mostra como a média dos valores reais (Paiva *et al.*, 2012) (EQUAÇÃO 5-9).

Equação 5: $$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (2.1)$$

Equação 6: $$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad (2.2)$$

Equação 7: $$MAPE = \frac{100}{n}\sum_{i=1}^{n}\frac{|y_i - \hat{y}_i}{y_i} \quad (2.3)$$

Equação 8: $$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (2.4)$$

Equação 9: $$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \quad (2.5)$$

3. METODOLOGIA

O Estudo discorre sobre a execução de experimentos ao longo da jornada exploratória para o processo de descoberta de conhecimento a partir de dados coletados, e os possíveis impactos na sociedade.

Para seu desenvolvimento, o estudo foi modelado como um problema de regressão em que a variável dependente (a ser predita) é o GPP e foi aplicada a técnica SHAP para avaliar o impacto de cada variável independente.

3.1 AQUISIÇÃO DO CONJUNTO DE DADOS

O universo de dados utilizado foi o do Bioma do Cerrado brasileiro, representado um conjunto de dados de dez estações meteorológicas de três estados e do DF, que compõem grande parte do cerrado: Brasília (DF), Goiânia (GO), Morrinhos (GO), Campo Grande (MS), Ponta Porã (MS), Três Lagoas (MS), Cuiabá (MT), Tangará da Serra (MT), Sorriso (MT), Campo Novo dos Parecis (MT) e Guarantã do Norte (MT). As estações escolhidas foram fundadas até dezembro de 2002, e apresentam dados abertos desde janeiro de 2003 a dezembro de 2020.

Foram incluídos dados abertos das dez estações meteorológicas coletadas da página da *internet do* INMET, de uma série histórica mensal de 2003 a 2020. Foram excluídos deste estudo os dados que não eram abertos, dados não oficiais, dados fora do intervalo da série histórica, dados das estações meteorológicas que possuem fundação posterior ao mês de janeiro de 2003, dados que representam outros biomas que não o do Cerrado, dados que não foram obtidos das plataformas e institucionais reconhecidas.

Para este estudo, foi realizado um recorte espacial do Bioma do Cerrado (Figura 8), utilizando o mapa vetorial do Bioma (mapa em escala de 1:250.000), disponibilizado pelo Instituto Brasileiro de Geografia e Estatística (IBGE), juntamente com o Ministério do Meio Ambiente e Mudança do Clima (MMA) , o qual serviu de insumo para capturar os dados de GPP, Temperatura da Superfície Terrestre, Fração Foto Absorvível, Índice de Área Foliar, Índice de Vegetação, de uma série histórica

mensal de Jan-2003 à dez-2020, disponíveis nos respectivos produtos de dados *MODIS*: MOD17A2 v006, MOD11A2 v061, MOD15A2 hv006 e MOD13A3 v006 (Kamel, 2015).

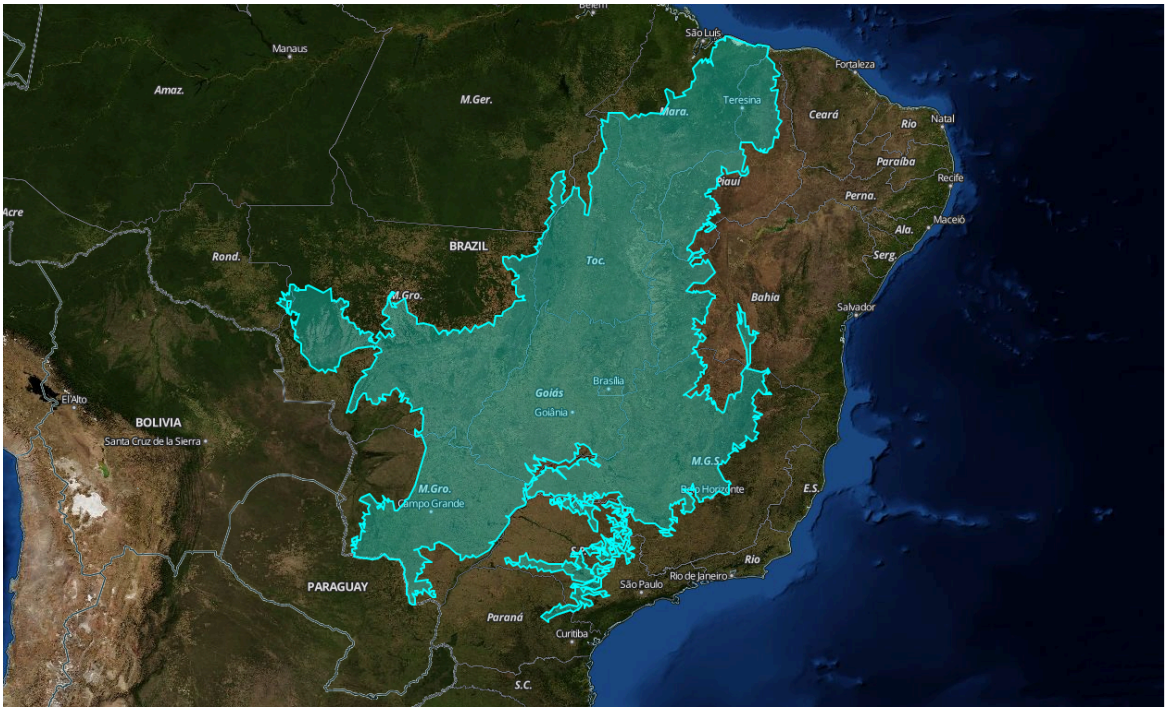

Figura 8. Mapa vetorial do bioma do cerrado
Fonte: Plataforma *Application for Extracting and Exploring Analysis Ready Samples* (AppEEARS)

O MODIS é o principal entre os cinco sensores dos satélite TERRA (Kamel, 2015; Wan, Hook e Hulley, 2021), lançado em 1999 e do AQUA, lançado em 2022, para medições do *Earth Observing System* (EOS), financiado pelo programa da NASA *Earth Science Enterprises* (ESE), programas de longa duração com fins de pesquisas de observação da superfície terrestre, oceanos e atmosfera, e suas interações mantido pelo comitê espacial Norte Americano (Anderson *et al*., 2022).

O MODIS foi projetado para satisfazer os requerimentos de três campos de estudos diferentes: atmosfera, oceano e terra, com bandas de resolução espectral e espacial selecionadas para o conhecimento de diferentes necessidades observacionais e para oferecer uma cobertura global quase diariamente, e possui algumas vantagens como: ampla cobertura espacial e espectral; continuidade nas tomadas de medidas nas regiões espectrais, já estimadas por outros satélites; usado para a meteorologia e monitoramento da temperatura da superfície do mar, gelo e vegetação; usado para monitorar a biomassa oceânica e os seus padrões de

circulação; e, é a primeira ferramenta dos satélites EOS na condução das pesquisas de mudanças globais.

A Figura 9 mostra a localização geoespacial de dez Estações Meteorológicas (Brasília, Goiânia, Morrinhos, Campo Grande, Ponta Porã, Três Lagoas, Cuiabá, Tangará da Serra, Sorriso, Campo Novo dos Parecis, Guarantã do Norte), através de coordenadas geográficas as quais são disponibilizados pelo Instituto Nacional de Meteorologia (INMET), baseados no critério de fundação das estações, de forma que essas variáveis capturadas representassem uma amostra do bioma do cerrado, a nível de: Precipitação Total Mensal, Pressão Atmosférica ao Nível Da Estação, Temperatura do Ar Bulbo Seco, Umidade Relativa do Ar, Vento Velocidade Horária, Radiação Global, variáveis que utilizamos para medir os impactos na previsão do GPP através dos valores SHAP.

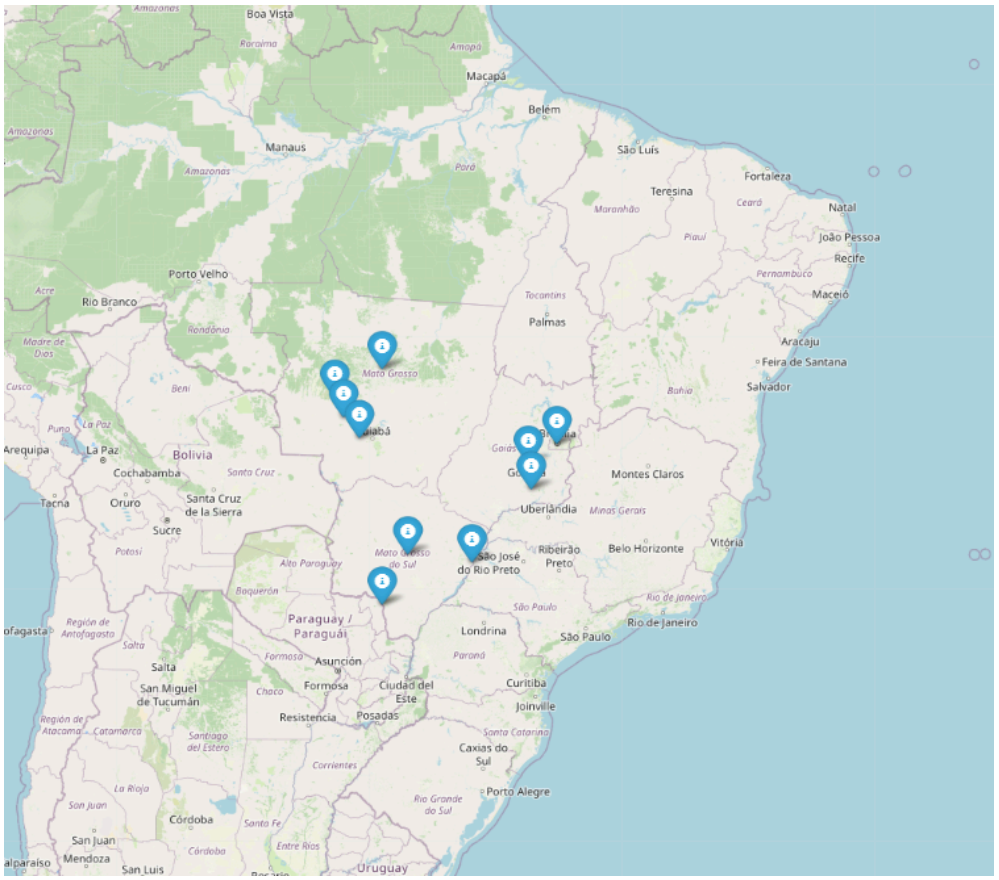

Figura 9. Mapa de localização com as respectivas latitudes e longitudes das estações meteorológicas do INMET, criada com a Biblioteca Folium em python

Os dados referentes às Áreas Queimadas e focos de incêndios se mostram relacionados ao mapa vetorial do bioma do Cerrado.

Os dados fornecidos pelo produto MOD17A2H Versão 6 (v061) que calcula a GPP, por meio de sensoriamento remoto, é um composto cumulativo de oito dias de valores com tamanho de pixel de 500 m com base no conceito de eficiência de uso de radiação que se apresenta potencialmente usado como entradas para modelos de dados para calcular fatores associados à produção de biomassa, sendo assim, o GPP, capturado pelo produto de dados, compreende-se como uma composição de diferentes variáveis conforme Equação 10.

Equação 10: $GPP = \epsilon_{max} m(T_{min})\, m(VPD)\, FPAR\, SW_{rad} 0.45$

Onde :

- ϵ_{max}: É a eficiência máxima do uso da LUE obtida a partir de uma tabela de referência, dependendo do tipo de vegetação, capacidade máxima da vegetação de converter radiação em biomassa.
- $m(T_{min})$: Um fator de escala que reduz a eficiência máxima (ϵ_{max}) em condições desfavoráveis de baixa temperatura mínima (T_{min}).
- $m(VPD)$: Um fator de escala que também reduz a eficiência máxima (ϵ_{max}) em condições de alto déficit de pressão de vapor (VPD), altos déficits de pressão de vapor resultam em fechamento estomático, limitando a absorção de CO^2.
- $FPAR$: Fração da radiação fotossinteticamente ativa absorvida pela vegetação, mede a quantidade de luz disponível que pode ser usada na fotossíntese.
- SW_{rad}: Radiação de onda curta (*Shortwave Radiation*), representa a energia solar disponível para o ecossistema.
- 0.45: Fator de conversão para ajustar as unidades e relacionar os valores de entrada ao resultado esperado em termos de produção de carbono.

3.1.1. Variáveis

A descrição das variáveis analisadas e as fontes constam do Apêndice 1.

Variáveis relativas ao Tempo e Data

- ANO – Identificação do ano das variáveis analisadas.
- MÊS – Identificação do mês das variáveis analisadas.

Variáveis relativas ao Clima

- GPP – Produtividade Primária Bruta acumulada mensalmente.
- TEMPERATURA DA SUPERFÍCIE– Temperatura média da superfície registrada no mês.
- PRECIPITAÇÃO TOTAL– Volume de chuva acumulado por mês, registrada por hora.
- PRESSÃO ATMOSFÉRICA AO NÍVEL DA ESTAÇÃO – Média da Pressão Atmosférica, registrada por hora.
- TEMPERATURA DO AR BULBO SECO – Temperatura do ar medida com termômetro seco, registrada por hora.
- UMIDADE RELATIVA DO AR – Média da porcentagem de umidade no ar registrada.
- VENTO VELOCIDADE HORÁRIA – Velocidade média do vento por hora.
- RADIAÇÃO GLOBAL – Intensidade da radiação solar incidente.

Variáveis relativas à Vegetação

- ÍNDICE DE VEGETAÇÃO – Indicador do vigor e densidade da cobertura vegetal.
- ÁREA DA FOLHA – Média da Extensão foliar calculada mensalmente, refletindo a densidade da vegetação.
- FRAÇÃO FOTO ABSORVÍVEL – Média da Proporção de luz absorvida pelas plantas no mês.

Variáveis relativas às Queimadas

- INCÊNDIOS – Número de focos de incêndio detectados na Região Centro-Oeste.
- ÁREA QUEIMADA CERRADO – Extensão da área afetada.

3.2 PRÉ-PROCESSAMENTO DOS DADOS

A etapa de pré-processamento e limpeza dos dados se mostra responsável pela redução de ruídos de diversos tipos que podem estar presentes nos *dataframes* analisados. Dados ruidosos são aqueles que contêm informações irrelevantes, inconsistências ou valores incorretos que podem distorcer análises e algoritmos de ML. Esses ruídos podem apresentar-se devido a erros de medição, problemas de transmissão, registro inadequado, anomalias naturais nos dados coletados, ou até mesmo indisponibilidade daquele dado. A implementação desta etapa se mostra fundamental para que se obtenha um bom resultado nos modelos de ML.

Durante o pré-processamento dos dados, sabe-se que a natureza das diferentes fontes, granularidades temporais, dados considerados com não números, pode gerar resultados inconsistentes, por isso, deve-se realizar manipulações para promover uma melhor integridade dos modelos testados. Para isso, foi utilizado manipulações estatísticas para promover a melhor qualidade dos dados como : a remoção de linhas e colunas inutilizadas, preenchimento de valores do tipo *Not a Number (NaN*) com 0, agrupando resultados dialisados com média ou mediana,

O processo incluiu a aquisição dos dados fornecidos pelo produto MOD17A2H Versão 6 (v061) que calcula a GPP, através de sensoriamento remoto, sendo esse um composto cumulativo de oito dias de valores com tamanho de pixel de 500 m com base no conceito de eficiência de uso de radiação que se apresenta potencialmente usado como entradas para modelos de dados para calcular fatores associados à produção de biomassa.

Destacamos alguns outros produtos de dados como o produto Terra *Moderate Resolution Imaging Spectroradiometer* (MODIS), *Land Surface Temperature/Emissivity 8-Day* (MOD11A2) Versão 6.1, fornecendo uma média de oito dias por *pixel de Land Surface Temperature and Emissivity (*LST&E), associado a temperatura da superfície terrestre, o produto combinado *Leaf Area Index* (LAI) *e Fraction of Photosynthetically Active Radiation* (FPAR) do *MOD15A2H* Versão 6.1, fornecendo dados de Índices de Áreas Foliares e fração fotossinteticamente ativa, assim como, dados de incêndios e Áreas queimadas, mapeados pelo Programa Queimadas do Instituto Nacional de Pesquisas Espaciais (INPE), na região

Centro-Oeste do país e dados do INMET coletados nas dez estações meteorológicas automáticas que se encontram englobadas no bioma do Cerrado, assim como na região Centro-Oeste, possuindo variáveis como Pressão Atmosférica e Temperatura do Ar no Bulbo Seco, entre outras.

3.3 ANÁLISE EXPLORATÓRIA

O pré-processamento, a análise exploratória inicial, o treinamento dos modelos de AM a partir dos dados gerados e o uso de *XAI,* pelo SHAP, se mostraram como fatores fundamentais para entender a análise de regressão dos dados do GPP, mensurando quais as variáveis mais importantes (Figura 10).

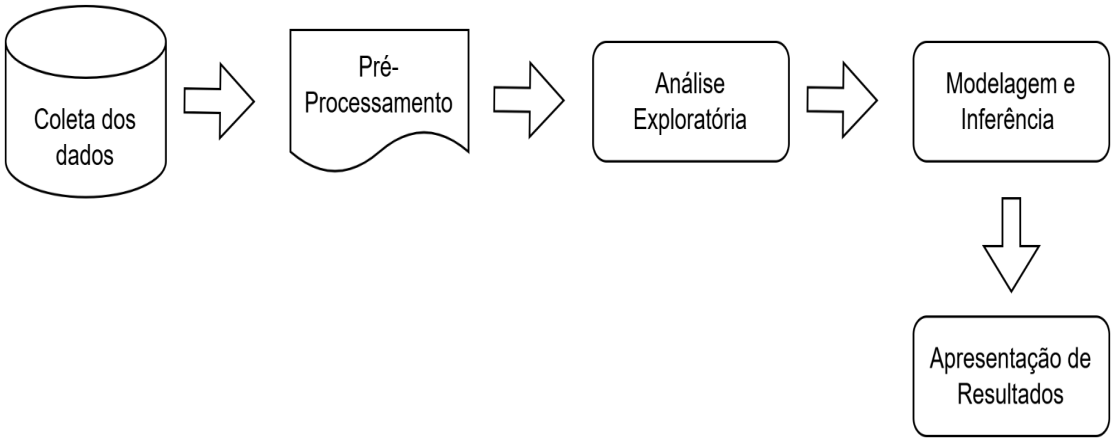

Figura 10. Etapas da pesquisa e Análise de dados
Fonte: Próprio do Autor

Nesta etapa foi observado o comportamento dos dados no que se refere à distribuição dos dados e das variáveis. Dessa forma, foram criadas diferentes visualizações dos dados de forma que tornasse mais tangível a leitura e entendimento dos mesmos.

3.4. MODELAGEM DOS DADOS

Para Modelagem de dados foram utilizadas diversas bibliotecas para a modelagem dos dados.

3.4.1 *Google Coolab e Bibliotecas Python*

O *Google Colab* (ou *Google Colaboratory*) é uma ferramenta gratuita baseada na nuvem, oferecida pelo Google, que permite escrever e executar código *Python* diretamente em um navegador da web. A ferramenta denota-se amplamente utilizada para ciência de dados, ML, IA e processamento de dados. No contexto deste projeto, a utilização do Google Colab mostrou-se fundamental, especialmente pela praticidade e portabilidade da infraestrutura oferecida pela plataforma.

Em conjunto com o *Google Colab*, o projeto se beneficiou do uso de várias bibliotecas *Python*, cada uma contribuindo com funções e recursos específicos para o desenvolvimento, implementação e análise dos resultados com técnicas XAI.

- **Pandas**: utilizada para manipular e analisar as tabelas. Essa biblioteca é atrelada a leitura, limpeza, transformação e organização dos dados para análises posteriores.
- ***Seaborn***: aplicada na criação de gráficos estatísticos. Permite criar perspectivas *datavis,* facilitando a análise de padrões e tendências.
- ***Matplotlib***: utilizada para criar gráficos personalizados, permitindo a construção de visualizações detalhadas e configuráveis, como histogramas, gráficos de barras e dispersão.
- ***SciPy***: aplicada para cálculos científicos e estatísticos avançados, incluindo otimização, integração, interpolação e testes estatísticos. Foi utilizada para validar o contexto de análise de distribuição das variáveis nos testes estatísticos paramétricos.
- **SHAP**: utilizada para interpretar modelos de ML, explicando o impacto de cada variável na previsão de modelos, ajudando na análise interpretativa; e para mensurar a importância das variáveis na análise de regressão do GPP.
- ***PyCaret***: empregada para simplificar processos de ML, incluindo pré-processamento, modelagem, seleção de parâmetros e teste de modelos. Foi utilizada para treinar e testar diferentes modelos de regressores disponibilizados.
- **Re (expressões regulares)**: aplicada para encontrar padrões em textos e manipular strings, facilitando a limpeza e organização dos dados.

- ***Folium***: usada para a criação de mapas interativos em Python, tornando a visualização de dados geoespaciais mais acessível e intuitiva
- ***Scikit-Learn* (*sklearn*)**: fornece ferramentas para ML, incluindo divisão de dados em treino e teste, parâmetros, teste cruzado e algoritmos de regressão e classificação. A sua utilização se mostrou fortemente atrelada às métricas de erro associadas aos valores reais comparados com os valores preditos do modelos treinados e testados pelo *PyCaret*.

O experimento configurado no *PyCaret* (Tabela 1) foi desenvolvido para resolver o problema de regressão, tendo como variável alvo (*target*) a coluna 'GPP', que representa uma variável do tipo contínua prevista.

Parâmetro	Valor
Session id	123
Target	GPP
Target type	Regression
Original data shape	(216, 13)
Transformed data shape	(208, 13)
Transformed train set shape	(143, 13)
Transformed test set shape	(65, 13)
Numeric features	12
Preprocess	TRUE
Imputation type	simple
Numeric imputation	mean
Categorical imputation	mode
Remove outliers	TRUE
Outliers threshold	0,0500
Normalize	TRUE
Normalize method	zscore
Fold Generator	KFold

Tabela 1. Parâmetros e Valores de Modelagem e Inferência

Para garantir a reprodutibilidade dos resultados, foi definido o parâmetro *session_id* = 123, fixando uma semente aleatória. A estratégia de teste cruzado escolhido foi o *K-Fold* com cinco divisões (*fold* = 5), permitindo dividir os dados em cinco partes, das quais quatro são utilizadas para treinamento e uma para teste,

repetindo o processo para cada *fold*. Essa abordagem reduz o risco de *overfitting* e melhora a generalização do modelo.

Além disso, foi ativada a remoção de *outliers* (*remove_outliers* = *True*) com um limiar de 0.05 (5%), eliminando valores extremos que poderiam distorcer os resultados. Também foi aplicada normalização (*normalize* = *True*) utilizando o método *Z-score*, padronizando os dados para que tivessem média zero e desvio padrão um, o qual se faz essencial para algoritmos sensíveis à escala. A divisão dos dados em treino e teste foi realizada de forma que 70% fossem destinados ao treinamento e 30% à teste. Com essa configuração, o experimento foi estruturado para analisar diferentes modelos e seus respectivos desempenhos.

Acerca do custo computacional, o PyCaret avaliou diferentes modelos regressores, porém o universo dos dados, do problema em vigor, é bem reduzido o que em termos de execução de cada algoritmo não envolveu um custo computacional elevado, tendo em média execuções inferiores a 0.30 segundos

3.4.2 ***Shapley Additive Explanations*** **(*SHAP*)**

O SHAP é uma forma de explicar qualquer resultado de ML. É uma abordagem teórica dos jogos para explicar a saída de qualquer modelo de ML. Ele conecta a alocação de crédito ideal com explicações locais usando os valores clássicos de Shapley da teoria dos jogos e suas extensões relacionadas.

Neste estudo o SHAP contribuiu para a análise das variáveis mais significativas no contexto da previsão da GPP.

3.5 APRESENTAÇÃO DOS RESULTADOS

Os resultados foram apresentados por meio de mapas de calor, gráficos de linhas, de cascata, histograma, de densidade e tabelas de contingência para identificar padrões temporais das variáveis analisadas.

3.6. CONSIDERAÇÕES ÉTICAS

Neste estudo todos os dados utilizados eram abertos. Dados secundários e abertos não envolvem interação direta com participantes de pesquisas, o que reduz as preocupações éticas relacionadas à privacidade, e confidencialidade e ao Consentimento Livre Informado (CLI). Entretanto, neste estudo zelou-se pela origem dos dados, pois somente foram utilizados dados de instituições oficiais de reconhecimento científico.

Este estudo também não contou com apoio financeiro e/ou institucional, sendo desenvolvido exclusivamente no âmbito do curso de Ciência da Computação da Universidade Federal de São Carlos (UFSCAR), cidade de São Carlos (SP).

4. RESULTADOS E DISCUSSÃO

A exploração do ambiente, o estilo de vida e comportamento humanos colocam em risco a vida no planeta, comprometem o ecossistema global e a biodiversidade (Brasil, 2020). Segundo Sawyer *et al*., (2017), o cerrado brasileiro é uma destas áreas ameaçadas. Desmatamento, mudanças climáticas, aquecimento global, agronegócio, fogo e queimadas, poluição e gases nocivos são fatores de risco (De Santana, Delgado e Schiavetti, 2020), e fizeram o mundo voltar o olhar para o 2^0 maior *hotspot* do Hemisfério Ocidental.

Tais problemas e desafios tendem a diminuir a *GGP*, uma parte essencial do ciclo da vida na terra (Pei *et al.,* 2022; Danelichen *et al*., 2015), e analisá-la é um imperativo. Segundo Zheng *et al*. (2020), isso não é uma tarefa fácil, uma vez que os modelos apresentam uma infinidade de dados, nem sempre são eficientes e muitos modelos de análise da GPP foram propostos (Pei *et al*., 2022; Zheng *et al*., 2020; Zhang *et al*., 2022; Li *et al*., 2022; Dias *et al*., 2024), e a GPP é influenciada por múltiplas variáveis (Li *et al*., 2022).

Nesse estudo a proposta consistiu em analisar a GPP por meio de modelos regressivos de ML, de forma que conseguíssemos entender o comportamento da variável alvo a ser predita por modelos como: RF Regressor, Adaboost Regressor, KNN, Linear Regression, entre outros. Apresentaremos os resultados desta análise nas três diferentes fases de execução deste estudo.

4.1 RESULTADOS DA ANÁLISE EXPLORATÓRIA

Durante o processo de Análise Exploratória inicial foi realizado o entendimento sobre as características principais, identificação de padrões, detecção de anomalias e verificação de relações entre variáveis envolvidas no nosso *dataset*.

A Figura 11, apresenta a evolução temporal da Fração Foto Absorvível e do GPP, destacando um padrão sazonal claro com variações cíclicas ao longo dos anos, sugerindo dependência com as estações do ano.

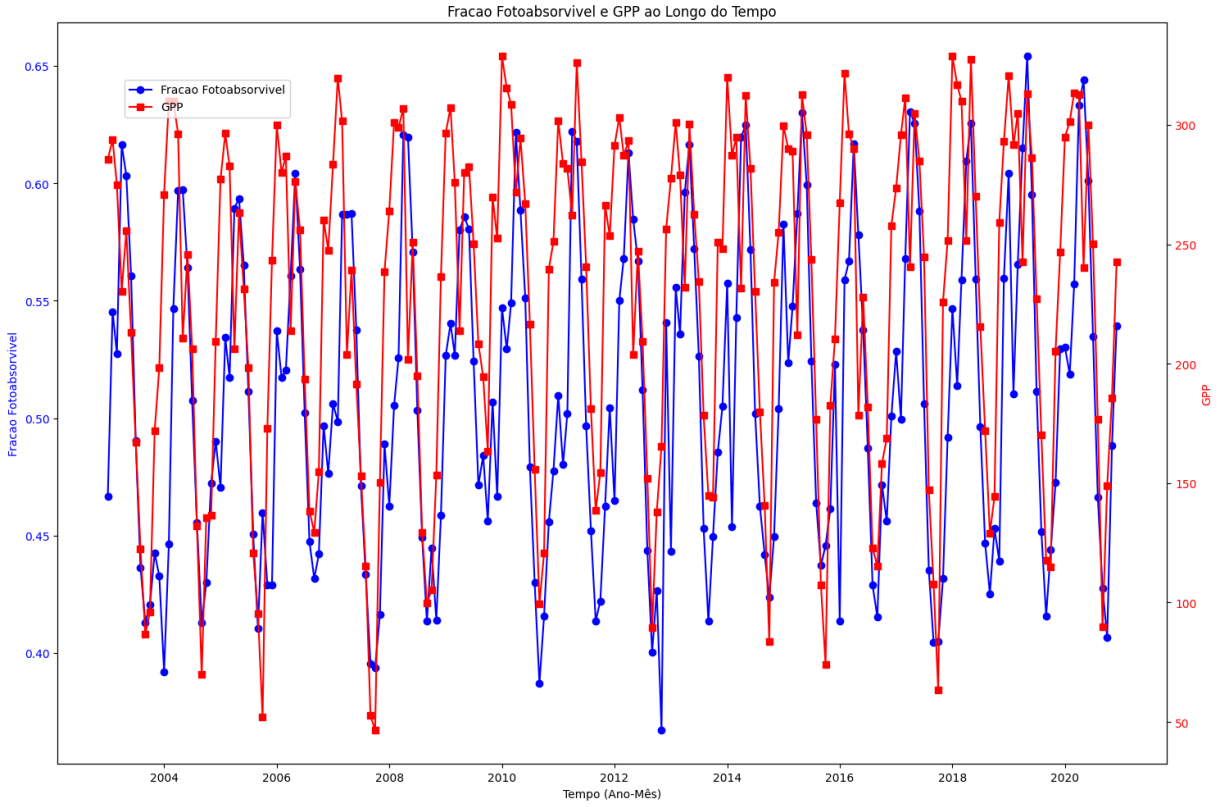

Figura 11. Fração Foto Absorvível e GPP ao longo do Tempo, Cerrado, Brasil, 2003-2020

Aparentemente, há uma correlação positiva entre as variáveis, conforme constatada pelo mapa de calor da Figura 14, no qual observa-se aumentos na Fração Foto Absorvível que acompanham aumentos no GPP, refletindo dependência direta da GPP da absorção de luz para a fotossíntese. As amplitudes das variações mostram possíveis influências climáticas ou ambientais em determinados anos, enquanto a ausência de tendências claras sugere estabilidade nos padrões sazonais ao longo do período analisado.

A Figura 12 apresenta a relação entre o GPP e a temperatura da superfície terrestre ao longo do tempo, revelando padrões interessantes. Ambos os conjuntos de dados mostram flutuações cíclicas, sugerindo uma forte sazonalidade. De fato, esse comportamento se mostra esperado, uma vez que o crescimento das plantas, representado pelo GPP, depende diretamente das condições climáticas, incluindo a temperatura.

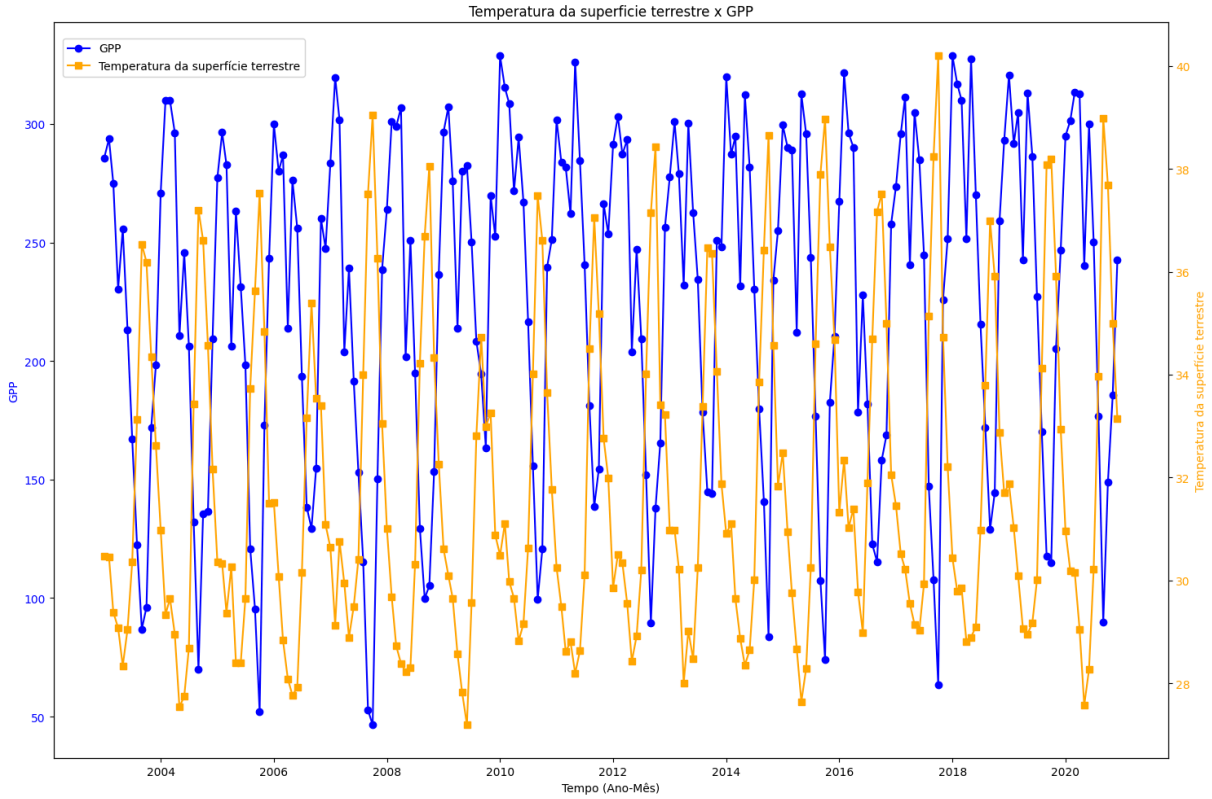

Figura 12. Análise da GPP relacionada à Temperatura da superfície, Cerrado, Brasil, 2003-2020

Observa-se que as variáveis estão inversamente correlacionadas, com os maiores valores de GPP ocorrendo em períodos de temperatura próximas a 28°C que podem favorecer a fotossíntese e o crescimento das plantas até certo ponto. No entanto, quando as temperaturas estão muito altas, o GPP também apresenta valores reduzidos, provavelmente devido à diminuição da atividade fotossintética durante estações mais quentes, o que pode indicar *estresse* térmico ou falta de água.

Outra análise peculiar se apresenta na Figura 13, sendo ela a evolução temporal dos Incêndios (azul) e da Área da Folha (verde) ao longo dos anos, evidenciando padrões sazonais com picos recorrentes em ambos os conjuntos de dados.

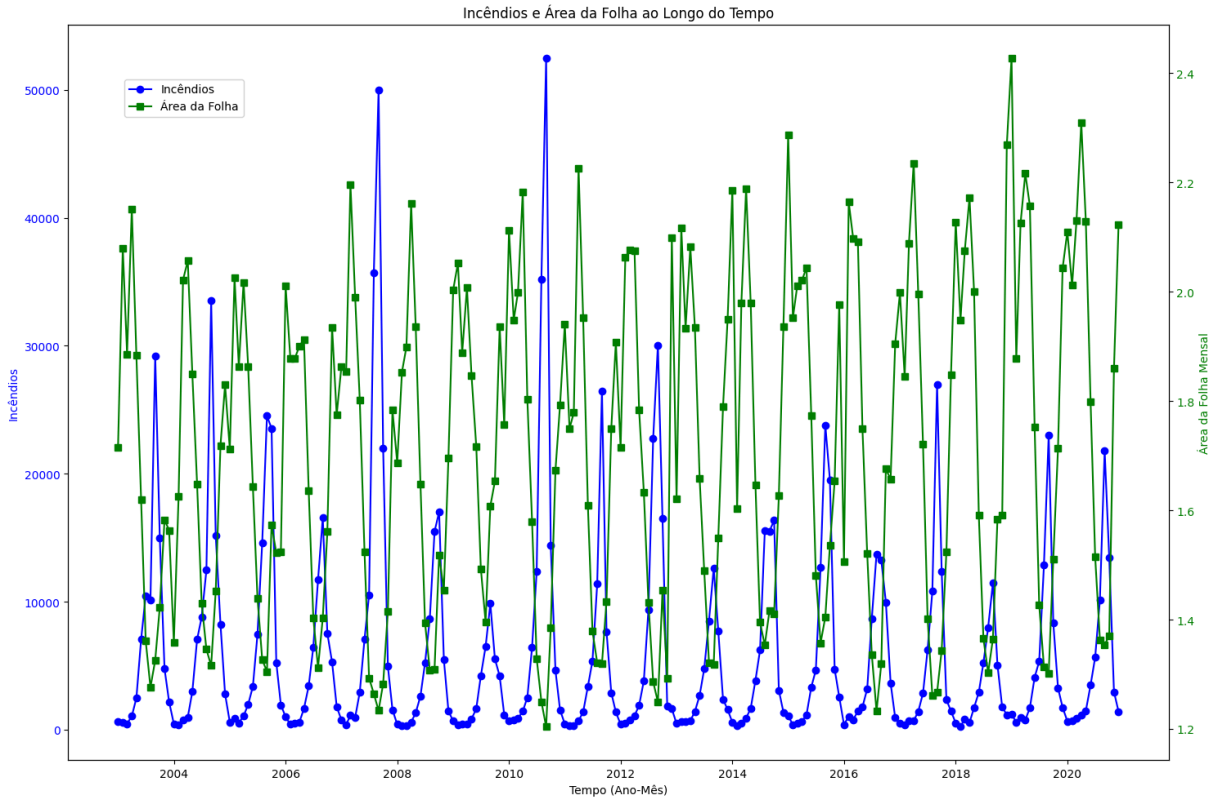

Figura 13. Frequência dos incêndios e área da folha ao longo do tempo, Cerrado, Brasil, 2003-2020

Os incêndios mostram aumentos bruscos, principalmente durante períodos específicos do ano, enquanto a Área da Folha exibe flutuações mais suaves e contínuas, possivelmente refletindo a recuperação gradual da vegetação após os eventos de incêndio. A redução na área da folha em períodos de alta incidência de incêndios indica perda significativa de biomassa, o que impacta diretamente a GPP. Isso ocorre porque a capacidade fotossintética das plantas, responsável pela absorção de luz e fixação de carbono, depende da área foliar disponível. Quando os incêndios reduzem drasticamente essa cobertura, a fotossíntese se mostra comprometida, levando a uma queda na produtividade do ecossistema e retardando sua recuperação ao longo do tempo.

Apresenta-se o Mapa de calor (Figura 14) que demonstra o coeficiente de correlação entre todas as variáveis numéricas presentes no dataframe.

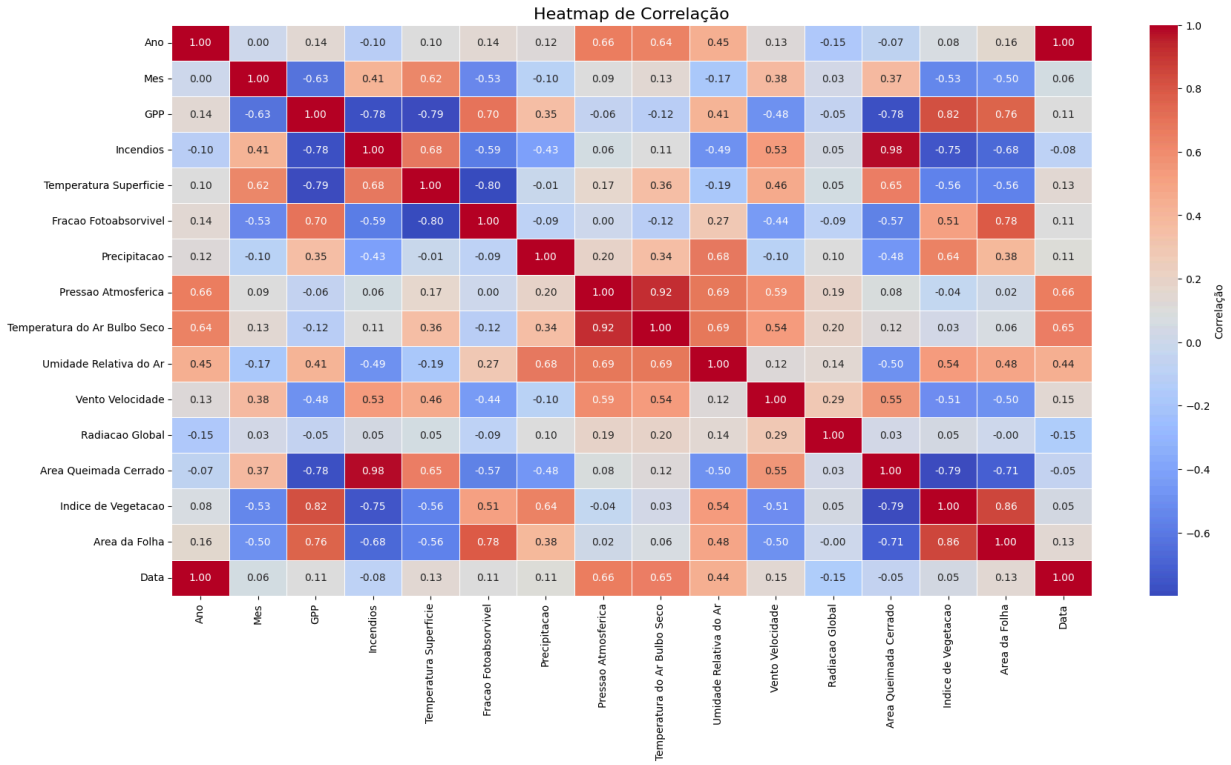

Figura 14. Análise de correlação da GPP e variáveis exploratórias, Cerrado, Brasil, 2003-2020

Percebe-se que as variáveis que mais possuem coeficiente de correlação positiva com a GPP (variável alvo) são: Área da Folha, Índice de Vegetação, Fração Foto Absorvível; enquanto que as variáveis que apresentaram coeficientes de correlações mais negativos com o GPP foram: Área Queimada Cerrado, Temperatura da Superfície e Incêndios.

4.2 DESEMPENHO DOS ALGORITMOS REGRESSORES (*PYCARET*)

Os resultados apresentados no *PyCaret* mostram que os modelos baseados em árvores de decisão e comitês de modelos (como ET e RF) obtiveram o melhor desempenho em termos de métricas de erro e poder preditivo (Tabela 2).

	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE
et	Extra Trees Regressor	23.8998	887.8515	29.7906	0.7891	0.1712	0.1353
rf	Random Forest Regressor	24.5891	920.2924	30.2568	0.7788	0.1742	0.1387
ada	AdaBoost Regressor	25.6323	956.9938	30.8968	0.7673	0.1745	0.1424
xgboost	Extreme Gradient Boosting	25.8172	1072.7904	32.6015	0.7465	0.1860	0.1455
llar	Lasso Least Angle Regression	26.3286	1045.4849	32.1261	0.7413	0.1891	0.1392
lasso	Lasso Regression	26.3284	1045.4177	32.1251	0.7413	0.1891	0.1392
gbr	Gradient Boosting Regressor	26.4132	1076.1750	32.5355	0.7335	0.1835	0.1467
ridge	Ridge Regression	26.3578	1083.0317	32.5859	0.7305	0.1996	0.1382
lr	Linear Regression	26.2867	1088.8706	32.6605	0.7284	0.1907	0.1380
br	Bayesian Ridge	26.9835	1111.7236	33.0608	0.7252	0.2437	0.1437
lightgbm	Light Gradient Boosting Machine	26.7985	1133.3665	33.5888	0.7212	0.1841	0.1463
en	Elastic Net	28.9495	1195.0394	34.4230	0.7110	0.2599	0.1587
huber	Huber Regressor	25.9567	1164.2107	33.6551	0.7090	0.2106	0.1401
par	Passive Aggressive Regressor	27.1265	1276.4821	34.8865	0.6902	0.1980	0.1485
knn	K Neighbors Regressor	29.4399	1333.3403	36.3387	0.6802	0.1953	0.1608
dt	Decision Tree Regressor	34.0001	1863.1569	43.1068	0.5489	0.2277	0.1882
omp	Orthogonal Matching Pursuit	38.4504	2641.6499	50.3357	0.3984	0.3118	0.2433
dummy	Dummy Regressor	58.6119	5009.8204	70.0996	-0.0798	0.3868	0.3783
lar	Least Angle Regression	69.1900	11612.2028	81.6962	-1.0864	0.4821	0.3843

Tabela 2. Erros dos modelos testados

Esses algoritmos, em particular o algoritmo ET apresentaram um desempenho sólido, destacando-se com os menores valores de erro entre os modelos testados. Seu MAE de 23.8998 indica que, em média, as previsões diferem cerca de 23.9 unidades dos valores reais, refletindo boa precisão geral. No entanto, erros maiores são penalizados no MSE de 887.8515 e no RMSE de 29.7906, sugerindo sensibilidade a valores extremos e *outliers*. O R^2 de 0.7891 mostra que o modelo explica aproximadamente 78.91% da variância nos dados, mas deixa cerca de 21% sem explicação, o que pode ser atribuído a ruídos, variáveis omitidas ou relações mais complexas entre os dados.

Adicionalmente, o RMSLE de 0.1712 indica que o modelo possui um bom desempenho em prever variações relativas, minimizando os efeitos de discrepâncias percentuais, enquanto o MAPE de 13.53% revela que os erros médios representam 13.53% do valor real, sendo aceitável para muitos casos práticos.

O RF apresentou um desempenho sólido, mas com erros ligeiramente superiores ao ET. O MAE foi de 24.5891, indicando que, em média, as previsões do

modelo diferiram cerca de 24.6 unidades dos valores reais. Esse valor é competitivo, mas demonstra uma leve perda de precisão em comparação com o ET, que teve um MAE de 23.8998. Já o MSE, que penaliza erros maiores, foi de 920.2924, resultando em um RMSE de 30.2568, mostrando que o modelo é sensível a *outliers* ou variações extremas nos dados, o que pode ter contribuído para uma menor precisão em casos específicos.

O R^2 do RF foi de 0.7788, explicando 77.88% da variância nos dados, um valor considerado bom, mas ligeiramente inferior ao 78.91% do ET, sugerindo que o modelo não capturou tão bem algumas variações complexas nas variáveis. O RMSLE de 0.1742 e o MAPE de 13.87% indicam que o modelo teve um bom desempenho para prever proporções relativas e diferenças percentuais, mas novamente ficou marginalmente atrás do ET. Esse comportamento pode ser explicado pela menor aleatoriedade nas divisões das árvores no *RF*, tornando-o ligeiramente mais propenso a *overfitting* em dados complexos ou com ruído.

O *AdaBoost Regressor* apresentou o terceiro melhor desempenho, sendo um desempenho moderado, com um MAE de 25.63 e RMSE de 30.89. Com um R^2 de 76.73%, ficou atrás dos modelos baseados em árvores como ET e RF. Além disso, o modelo apresentou um MAPE de 14.24%, reforçando sua vulnerabilidade a ruídos nos dados e possíveis *overfittings* devido ao peso dado a erros em cada iteração.

Um dos modelos que apresentou desempenho ruim foi o *Dummy Regressor*. Sendo ele um modelo simples utilizado principalmente como *baseline* para comparar o desempenho de modelos preditivos mais complexos, que não realiza aprendizado real ou análise de padrões nos dados. Em vez disso, faz previsões baseadas em estratégias triviais, como média, mediana, valor constante ou último valor observado no conjunto de dados de treinamento. Seu principal objetivo é fornecer uma referência mínima para determinar se um modelo mais sofisticado realmente melhora a previsão em relação a abordagens aleatórias ou simplistas.

O *Dummy Regressor* obteve um R^2 de -0.0798, indicando que ele não conseguiu explicar nenhuma variância nos dados e, na verdade, performou pior do que simplesmente prever a média dos valores observados. Os dados possuem padrões que não podem ser capturados por estratégias simples e exigem modelos mais complexos para serem devidamente explicados.

4.3 ANÁLISE DE IMPORTÂNCIA DE VARIÁVEIS E VALORES SHAP

A importância das variáveis (*Feature Importance*) é uma técnica utilizada em ML para medir a relevância ou contribuição de cada variável de entrada (features) em um modelo preditivo. Ela indica o peso ou influência de cada variável na tomada de decisão do modelo, a análise de importância das variáveis para os modelos ET, RF, Adaboost foi extraída pela biblioteca do Pycaret, posteriormente analisamos separadamente cada modelo citado anteriormente, fora do ambiente de configuração do Pycaret com o uso do SHAP, para explicar quanto cada variável contribui para a previsão do modelo.

Os valores de *Feature Importance*, extraído com Pycaret, no modelo ET, representam a contribuição relativa de cada variável na construção do modelo preditivo, variando entre 0 e 1, com a soma totalizando 1 (ou 100%) (Figura 15).

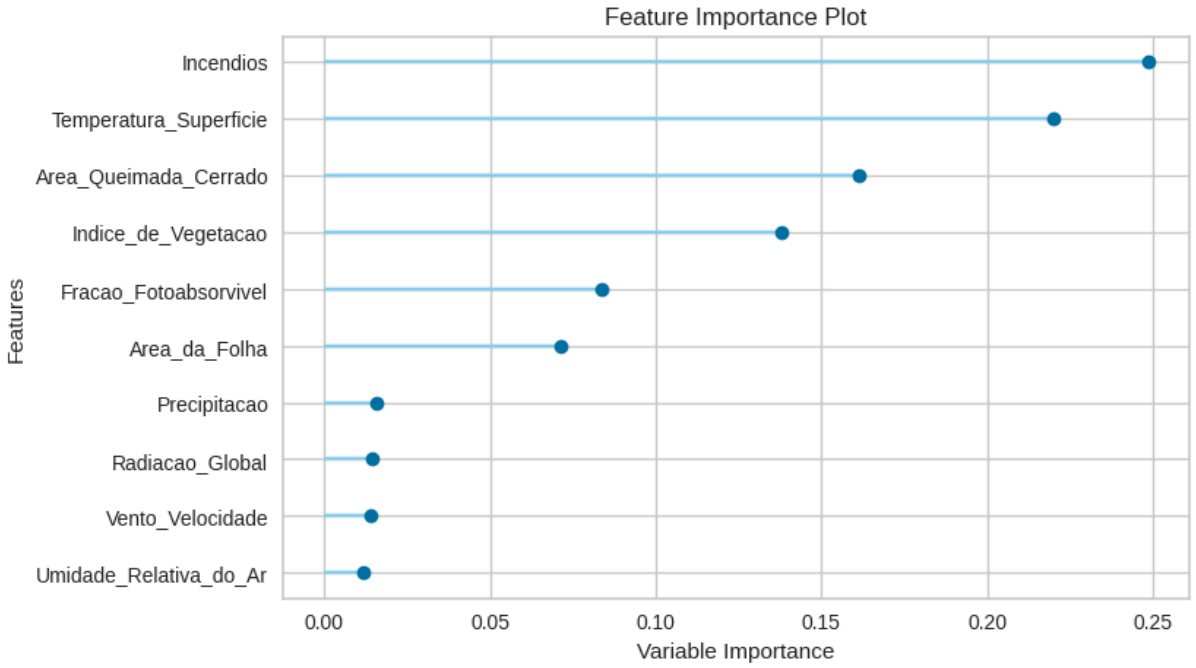

Figura 15. *Feature Importance* do Modelo ET, Cerrado, 2003-2020

As variáveis Incêndios (0.25) e Temperatura_Superfície (0.22) se destacam como as mais influentes, sendo responsáveis por 47% da importância total, o que indica forte impacto na previsão do modelo. Em seguida, variáveis como Área_Queimada_Cerrado (0.15) e Índice_de_Vegetacao (0.14) possuem valores intermediários, contribuindo de forma significativa, mas com menor peso individual. Já variáveis como Fracao_Fotoabsorvivel (0.10) e Área_da_Folha (0.08) apresentam

relevância moderada, enquanto Precipitacao, Radiação_Global, Vento_Velocidade e Umidade_Relativa_do_Ar possuem valores abaixo de 0.05, indicando impacto limitado na previsão.

Os modelos RF e ET, embora baseados em árvores de decisão e pertencentes à mesma classe de algoritmos *ensemble*, diferem significativamente na forma como constroem suas árvores e atribuem importância às variáveis preditoras, resultando em comportamentos distintos na modelagem.

O RF seleciona os pontos de divisão (*splits*) de forma ótima, buscando maximizar a redução da impureza (ou variância) em cada nó. Esse processo favorece variáveis com padrões claros e dominantes, resultando em uma concentração maior de importância em poucas variáveis-chave. Por outro lado, o ET realiza as divisões de forma aleatória dentro dos limites dos dados, sem buscar a otimização. Essa abordagem mais aleatória reduz a dependência de variáveis específicas e distribui a importância de maneira mais uniforme, tornando o modelo mais robusto contra variações nos dados e menos sensível a *outliers*.

No RF, as variáveis mais fortemente correlacionadas com o alvo recebem pesos mais altos. As variáveis Incêndios (48%) e Temperatura_Superfície (21%) somam 69% da importância total, refletindo a tendência do RF de priorizar estabilidade e previsibilidade ao concentrar o impacto em poucas variáveis dominantes. Esse comportamento é considerado conservador, pois o modelo foca em padrões claros e previsíveis.

Por outro lado, no ET, a importância se apresenta distribuída de forma mais democrática. Incêndios (25%) e Temperatura_Superfície (22%) somaram 47%, permitindo maior influência de variáveis secundárias. Esse comportamento exploratório torna o ET mais adequado para capturar padrões complexos e não lineares, especialmente em dados com alta dimensionalidade ou ruído.

Na Figura 16 está demonstrado o *Feature Importance* , extraído pelo Pycaret, do modelo RF destacando as variáveis mais relevantes para a previsão do alvo.

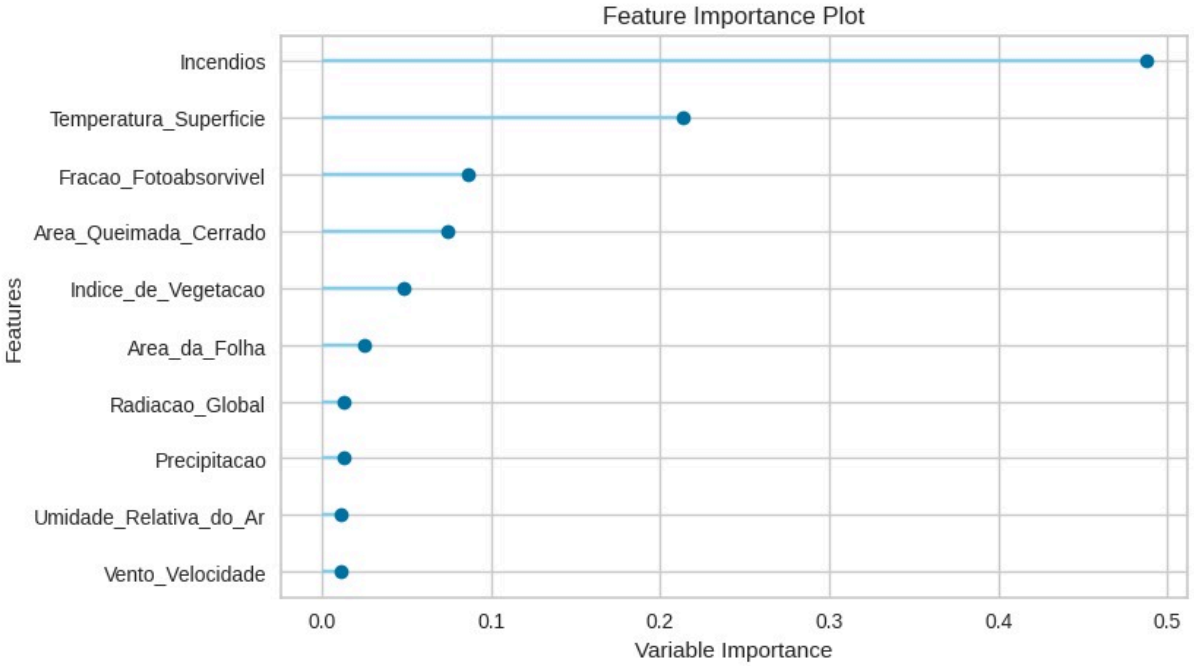

Figura 16. *Feature Importance* do Modelo RF, Cerrado, Brasil, 2003-2020

Incêndios foi a variável de maior influência, se destacando como a variável mais importante; a Temperatura_Superfície, se destaca sugerindo forte influência na previsão; variáveis como: Fracao_Fotoabsorvivel, Área_queimada e Índice_de_Vegetacao também têm influência , mas menor do que as duas principais, enquanto variáveis como Área_da_Folha, Precipitacao, Radiacao_Global, Umidade_Relativa_Ar, Pressao_Atmosferica e Vento_velocidade apresentam pouca ou nenhuma influência.

Na Figura 17 está demonstrado a *Feature Importance* do Modelo *Adaboost Regressor*, extraída do Pycaret.

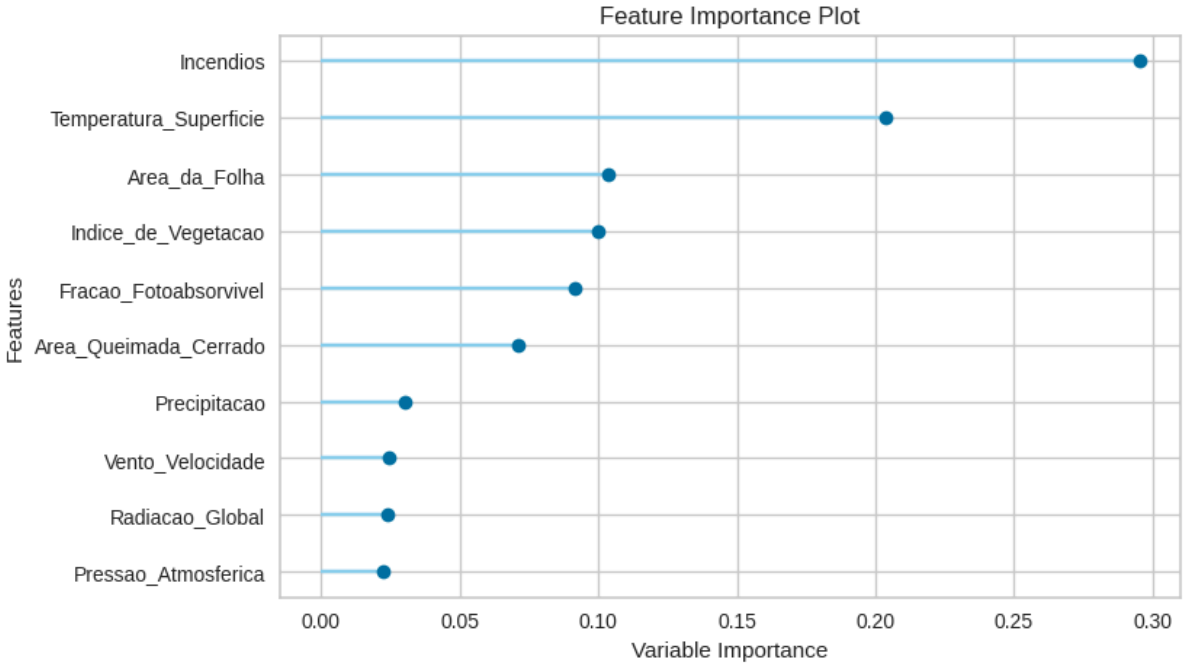

Figura 17. *Feature Importance* do Modelo *Adaboost Regressor*, Cerrado, Brasil, 2003-2020

Incêndios e Temperatura_Superfície se destacam como as variáveis mais impactantes, com pesos aproximados de 0.30 e 0.22, respectivamente, sugerindo forte influência na previsão; variáveis como Área_da_Folha, Índice_de_Vegetacao, Fracao_Fotoabsorvivel e Área_queimada também têm influência significativa, mas menor do que as duas principais, enquanto variáveis como Precipitacao, Vento_velocidade, Pressao_Atmosferica e Radiacao_Global apresentam menor importância.

Após analisarmos as métricas de erros em relação aos valores reais e os valores previstos pelos modelos testados no Pycaret, optamos por analisar os valores SHAP fora do ambiente de configuração do mesmo do Pycaret, de forma que pudesse ter uma análise mais detalhada dos modelos selecionados. A importância média absoluta dos valores SHAP se apresenta como uma métrica usada para quantificar a relevância de cada variável em um modelo de aprendizado de máquina com base nos valores SHAP. Podemos mensurar a relevância de cada variável no modelo ET (Figura 18), destacando o impacto de cada uma das previsões. No eixo X, a magnitude média do impacto (mean(|SHAP value|)) mostra a influência relativa das variáveis, enquanto no eixo Y elas estão ordenadas da mais importante para a menos importante.

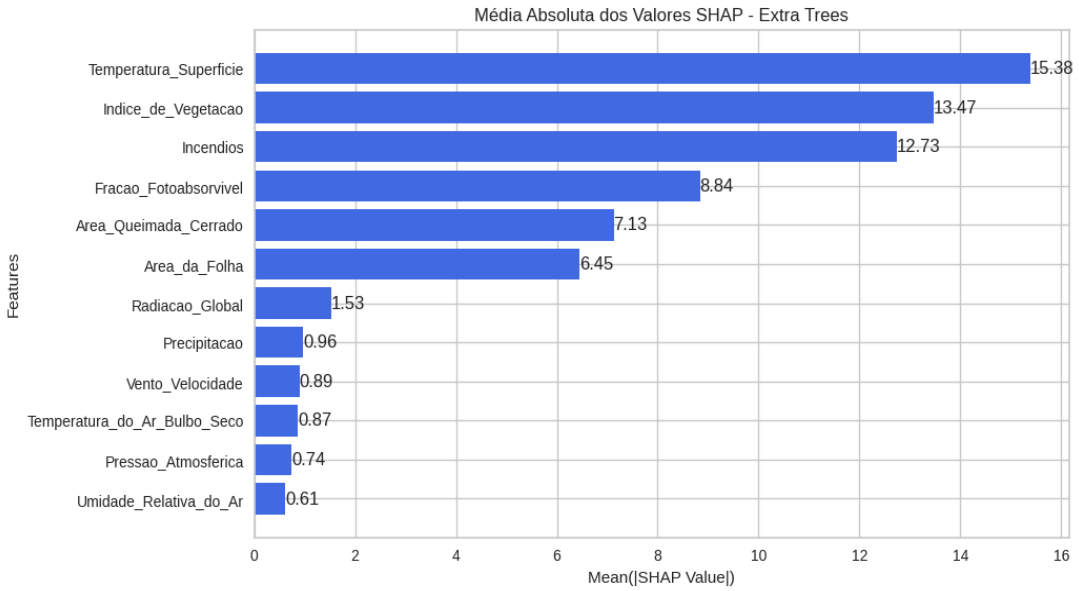

Figura 18. Média Absoluta dos Valores SHAP- *ET,* Cerrado, Brasil, 2003-2020

A média absoluta dos valores SHAP das variáveis no modelo Extra Trees, levando em consideração todos os *folds* da teste cruzado. A variável Temperatura_Superfície teve o maior impacto nas previsões do modelo, com um valor médio SHAP de 15.38, seguida por Índice_de_Vegetação (13.47) e Incêndios (12.73), indicando que essas três variáveis foram as mais influentes na tomada de decisão do modelo. Outras variáveis com impacto relevante incluem Fração_Fotoabsorvível (8.84), Área_Queimada_Cerrado (7.13) e Área_da_Folha (6.45), sugerindo que características relacionadas ao ambiente e vegetação desempenham um papel importante na modelagem. Em contrapartida, variáveis como Radiação_Global (1.53), Precipitação (0.96), Vento_Velocidade (0.89) e Temperatura_do_Ar_Bulbo_Seco (0.87) tiveram menor impacto, indicando uma influência reduzida nas previsões do modelo. As variáveis menos relevantes, como Pressão_Atmosférica (0.74) e Umidade_Relativa_do_Ar (0.61), apresentam os menores valores SHAP médios, sugerindo que possuem pouca relevância no desempenho do modelo.

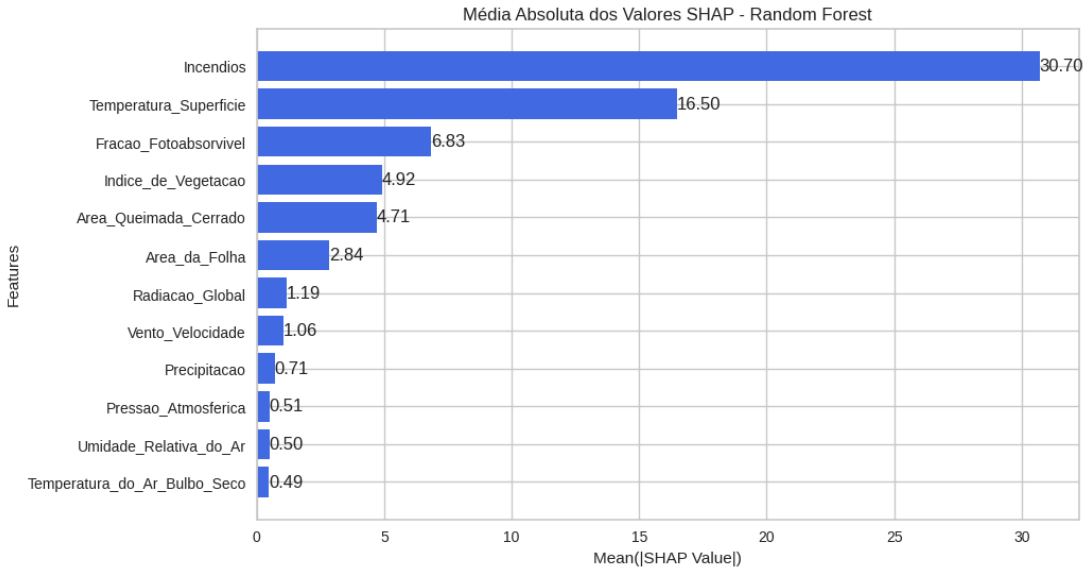

Figura 19. Média Absoluta dos Valores SHAP- *RF,*Cerrado, Brasil, 2003-2020

A Figura 19 exibe a média absoluta dos valores SHAP das features no modelo RF, considerando todos os *folds* da teste cruzado. A variável Incêndios apresentou o maior valor SHAP médio (30.70), indicando que teve a influência mais significativa na predição. Em seguida, Temperatura_Superfície (16.50) também se destacou como uma feature crucial. Outras variáveis, como Fração_Fotoabsorvível (6.83), Índice_de_Vegetação (4.92) e Área_Queimada_Cerrado (4.71), tiveram impacto relevante, mas menor em comparação às duas primeiras. Por outro lado, variáveis como Pressão_Atmosférica (0.51), Umidade_Relativa_do_Ar (0.50) e Temperatura_do_Ar_Bulbo_Seco (0.49) apresentaram baixa importância, sugerindo que têm pouca influência nas previsões do modelo.

A distribuição mais homogênea dos valores médios absolutos SHAP para as variáveis no *ET* pode ser justificada principalmente pela aleatoriedade no critério de divisão dos nós, que reduz a concentração da importância em poucas variáveis e favorece uma distribuição mais ampla da relevância entre diferentes features. Enquanto o *RF* escolhe os pontos de divisão com base no melhor critério (maximização da redução de impureza, como Gini ou entropia), o ET seleciona aleatoriamente os pontos de corte dentro das features disponíveis

O Gráfico Beeswarm, se mostra como uma ferramenta visual amplamente utilizada para interpretar modelos de aprendizado de máquina, especialmente em conjunto com os valores SHAP. Ele apresenta a distribuição dos impactos de cada variável sobre as previsões, permitindo uma análise global do comportamento do

modelo. No eixo Y, denota-se listadas as variáveis ordenadas por importância, enquanto o eixo X mostra a magnitude e o sinal (positivo ou negativo) do impacto das variáveis. Valores SHAP positivos indicam que a variável aumenta a previsão, enquanto valores negativos sugerem que ela reduz o valor previsto. Cada ponto no gráfico representa um exemplo (instância) do conjunto de dados, proporcionando uma visão detalhada de como as variáveis influenciam os resultados.

As cores dos pontos indicam os valores reais das variáveis em cada instância, geralmente com tons mais claros (vermelho) representando valores altos e tons mais escuros (azul) representando valores baixos. Essa coloração facilita a identificação de correlações entre os valores das variáveis e o impacto nas previsões. Diante do cenário de teste cruzado com um K-fold = 5, selecionamos o fold 2 para analisar como os modelos *ET*, e *RF* se comportam em relação à distribuição dos impactos das variáveis, permitindo uma análise do comportamento específico do modelos em partes dos conjunto de dados, refletindo como as variáveis influenciam as previsões nesse subconjunto.

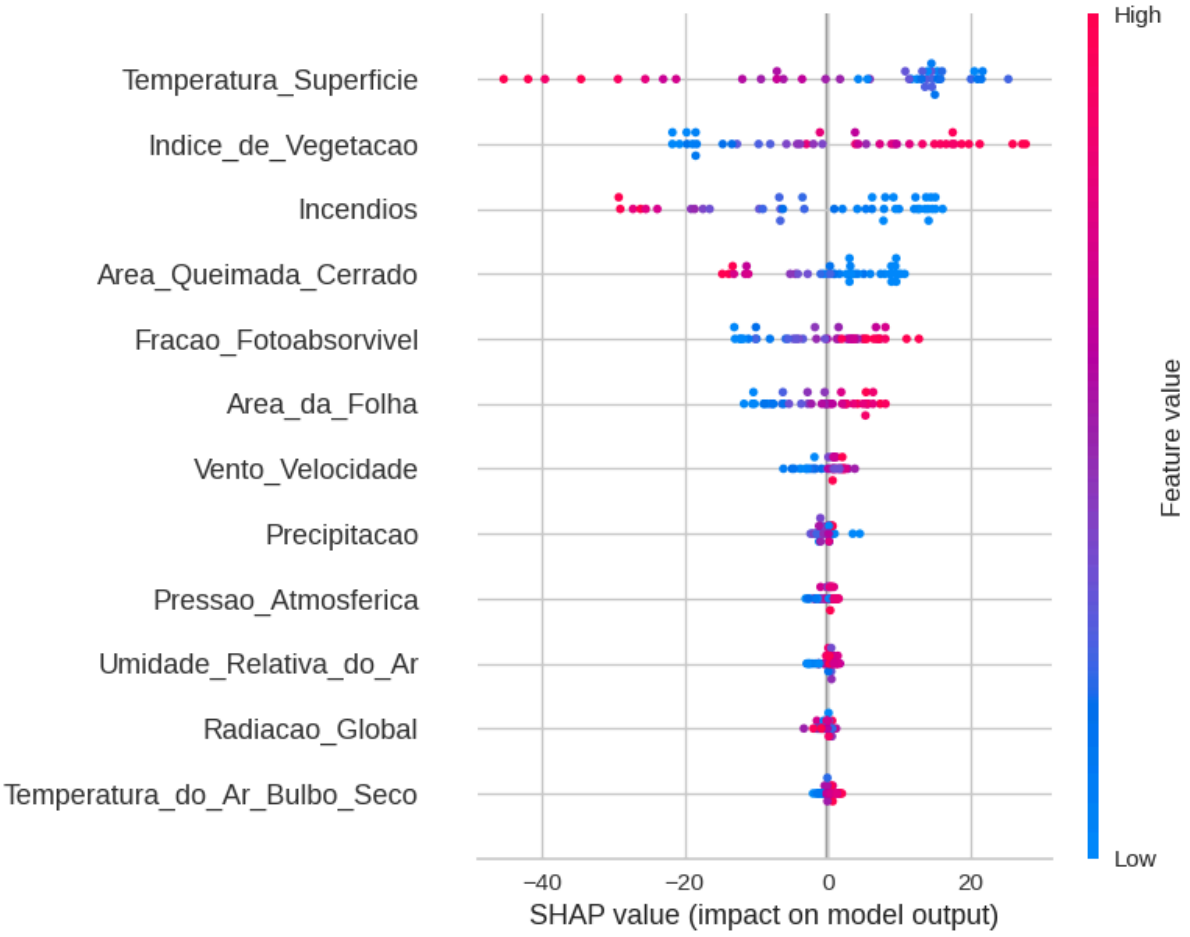

Figura 20. Gráfico de Beeswarm para o modelo ET, Cerrado, Brasil, 2003-2020

A Figura 20, mostra o gráfico de Beeswarm, para o *fold* 2, do modelo ET, onde a variável Temperatura_Superfície apresentou o maior impacto no modelo, com valores elevados reduzindo significativamente as previsões, enquanto valores mais baixos tiveram um efeito mais neutro. A variável Índice_de_Vegetação também exerceu influência expressiva, com valores altos associados a previsões a impactos positivos nas previsões, demonstrando sua relevância. Da mesma forma, Incêndios se destacou, apresentando uma distribuição de impactos onde valores altos geram impacto negativo no valor predito pelo modelo.

Variáveis de impacto moderado, como Fração_Fotoabsorvível, mostraram efeitos positivos para valores mais altos, sugerindo uma relação relevante, porém menos dominante. Outras variáveis como Área_da_Folha e Vento_Velocidade também demonstraram impactos menores, mas ainda perceptíveis, indicando uma influência complementar às principais variáveis.

Por outro lado, variáveis como Radiação_Global, Precipitação, Pressão_Atmosférica e Umidade_Relativa_do_Ar exibiram impactos reduzidos ou quase nulos, sugerindo que, embora processadas pelo modelo, suas contribuições para a previsão são secundárias em comparação com as variáveis de maior impacto.

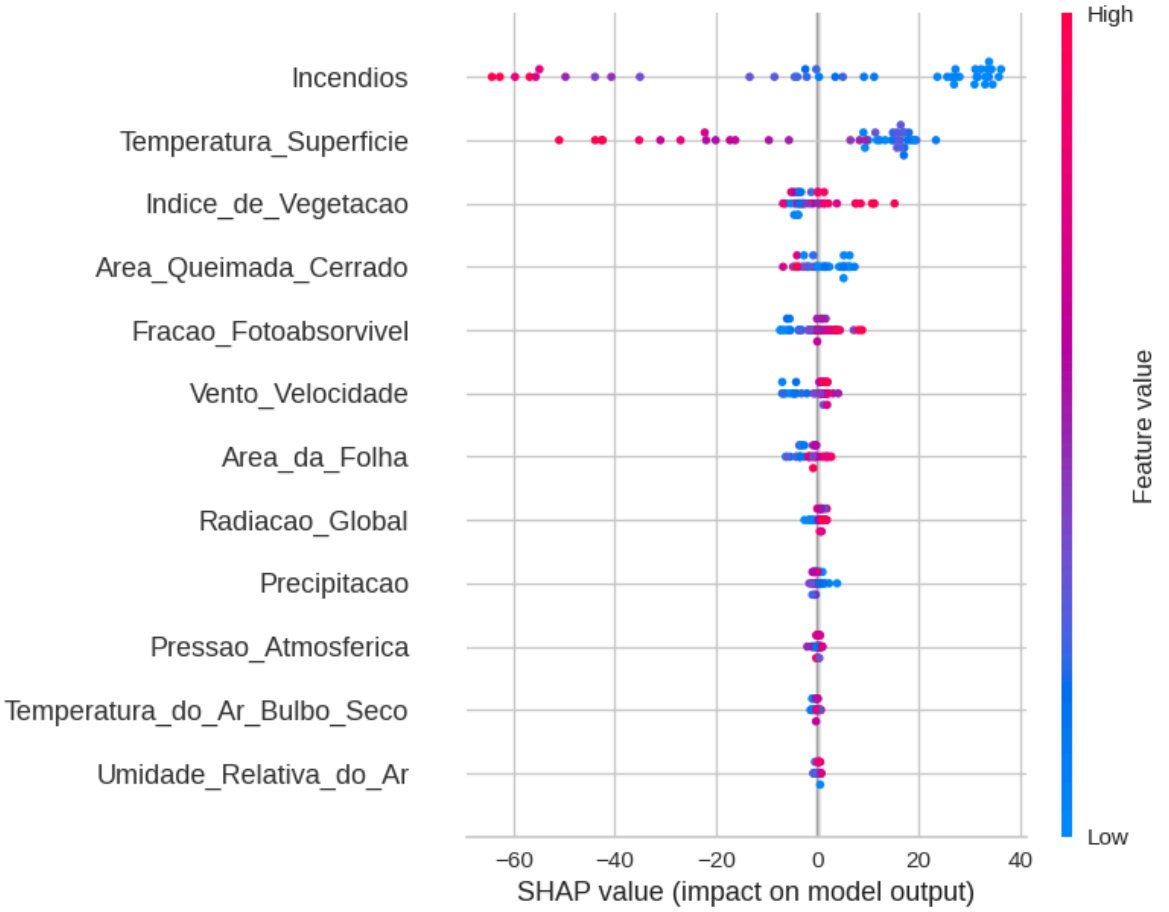

Figura 21. Gráfico de Beeswarm para o modelo ET, Cerrado, Brasil, 2003-2020

Para o modelo RF (Figura 21), entre as variáveis, a de maior relevância se mostra como a variável Incêndios, sendo a mesma a de maior expressividade no topo do gráfico, para o *fold* 2, o modelo interpreta dados com maior quantidade de incêndios reduzindo os valores da previsão. Altos valores dessa variável, em vermelho, tendem a ter impactos negativos expressivos, reduzindo as previsões, enquanto valores mais baixos (azul) apresentam impacto positivo. Isso sugere que um maior número de incêndios está associado à redução nas previsões feitas pelo modelo.

A variável Temperatura_Superfície também se destaca, sendo a segunda mais importante. Altos valores dessa variável estão ligados a impactos negativos, o que implica que temperaturas mais elevadas tendem a diminuir as previsões. Já valores mais baixos provocam um leve aumento nas previsões.

Ainda sobre essa ótica variável Indice_de_Vegetacao possui um impacto menor em relação às anteriores. Altos valores dessa variável aumentam levemente as previsões, enquanto valores mais baixos possuem uma densidade muito próxima

de valores nulos. Isso pode indicar que áreas com maior capacidade de absorção de luz, sugerindo maior biomassa, estão associadas a riscos específicos analisados pelo modelo. Por outro lado, a variável Área_Queimada_Cerrado apresenta influência moderada. Valores mais baixos tendem a aumentar as previsões, sugerindo que áreas previamente queimadas contribuem para previsões mais altas.

O Gráfico waterfall se mostra como uma ferramenta visual amplamente utilizada para explicar como diferentes variáveis possuem um impacto local a nível de uma amostra, ao contrário do gráfico de Valores Médios Absolutos Shap e Beeswarm que trazem avaliações globais, podemos assim entender como os valores SHAP se apresentam para uma previsão de um exemplo em específico. Ele detalha a decomposição do valor previsto, partindo de um valor base, também conhecido como valor esperado ($E[f(X)]$), que representa a média das previsões para todos os dados do conjunto de teste do *fold* analisado. Primeiramente, o gráfico exibe as contribuições individuais de cada variável, ilustradas por barras coloridas. Barras vermelhas indicam impactos positivos, aumentando a previsão, enquanto barras azuis representam impactos negativos, reduzindo o valor previsto. Essas contribuições são somadas incrementalmente ao valor base, resultando na previsão final ($f(x)$) do modelo para a amostra analisada. Esse formato é especialmente útil para explicar previsões individuais, destacando quais variáveis tiveram maior influência nos resultados e fornecendo transparência ao comportamento do modelo. Além disso, permite identificar fatores-chave e diagnosticar possíveis ajustes necessários.

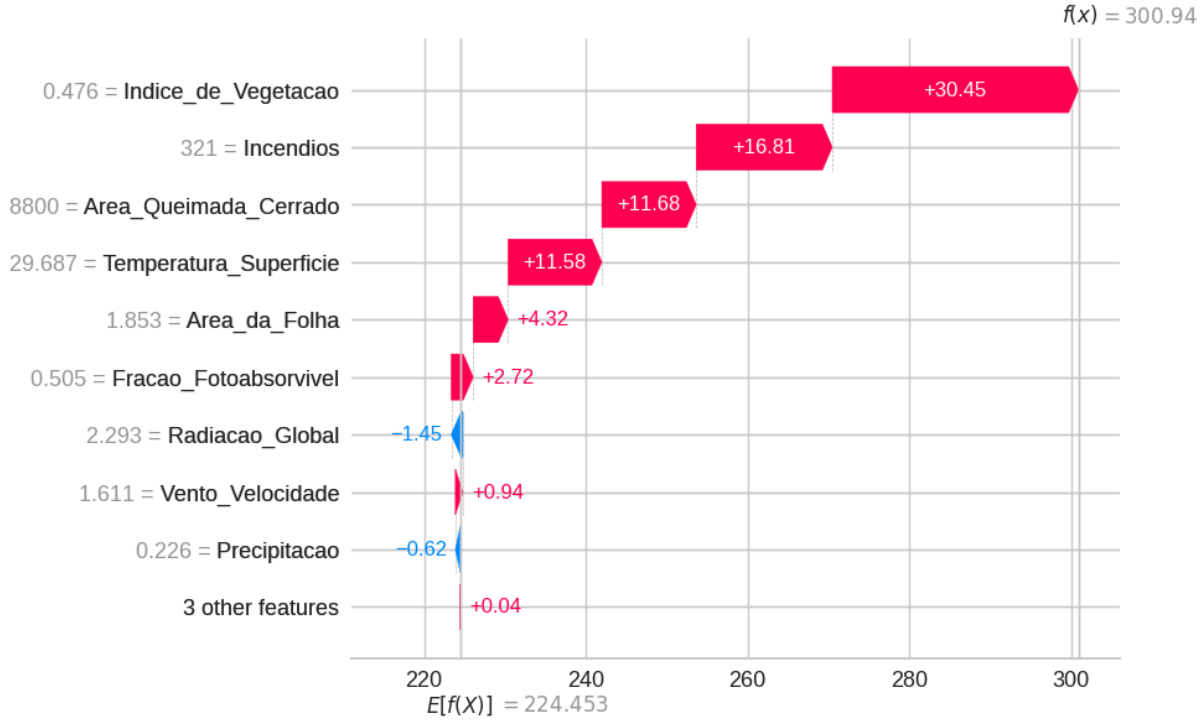

Figura 22. *Waterfall Plot* para o modelo ET, Cerrado, Brasil, 2003-2020

O Gráfico W*aterfall* para o modelo ET (Figura 22) mostra a decomposição da previsão para uma amostra específica, selecionamos a amostra 17, referente ao *fold* 2, começando com o valor base esperado $(E[f(X)])$ de 230.733 e ajustando até o valor final previsto $(f(x))$ de 150.708. Ele detalha como cada variável contribui para essa previsão, com barras azuis indicando reduções e barras vermelhas representando aumentos.

A variável Índice_de_Vegetação teve o maior impacto positivo, aumentando a previsão em +30.45, e isso indica que para essa amostra em específico o Indice de Vegetação influenciou positivamente o valor final previsto. A variável Incêndios também apresentou impacto positivo significativo, aumentando em +16.81. A Área_Queimada_Cerrado (+11,68), Temperatura_Superficie (+11,58) e Area_da_folha (+4,32) também contribuíram para um aumento positivo no valor previsto final. Por outro lado, as variáveis Fração_Fotoabsorvível (+2,72), Radiação_Global (-1.45), Vento_Velocidade (+0.94), Precipitacao (-0,62), tiveram menores impactos na previsão final.

O Gráfico waterfall do modelo RF (Figura 23) apresenta a decomposição da previsão para uma amostra específica, partindo de um valor base esperado ($E[f(X)]$) de 224.241 e chegando ao valor final previsto ($f(x)$) de 298.141. Cada variável contribui positiva ou negativamente para essa previsão, com as barras azuis indicando reduções e as vermelhas aumentando o valor previsto.

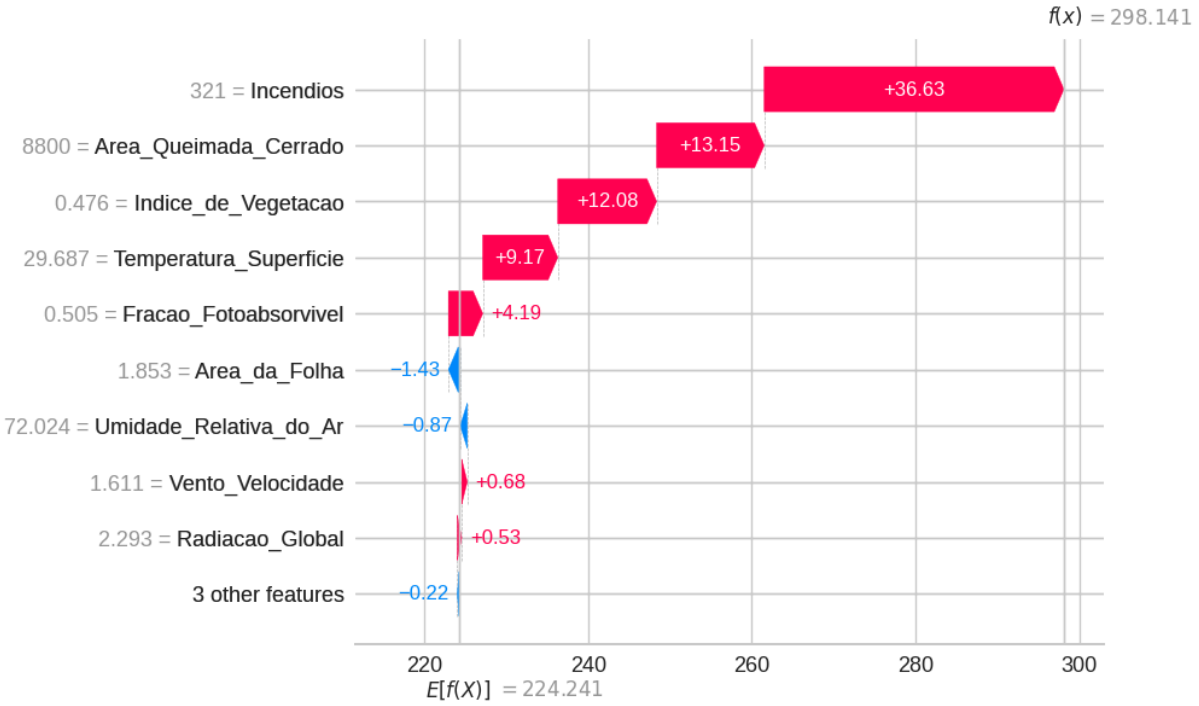

Figura 23. *Waterfall Plot* para o modelo RF, Cerrado, Brasil, 2003-2020

A variável Incêndios teve o maior positivo, para a mesma amostra 17 analisada em específica, aumentando o valor em +36.63. A segunda variável de maior impacto foi Area_Queimada_Cerrado também apresentou um aumento expressivo de +13.15. As variáveis Indice_de_Vegetacao (+12,08), Temperatura_Superficie (+9,17) Fracao_Fotoabsorvivel (+4,19), impactaram positivamente o valor previsto final. Já as variáveis restantes tiveram impactos de menor significância para o valor previsto final, com valores inferiores a (± 2,00).

Logo, percebe-se que as visualizações de Valores Médios Absolutos SHAP, Beeswarm e Waterfall, quando combinadas, conseguem traduzir uma visualização a nível global e local dos conjuntos de dados analisados. Assim como elas, existem outros tipos de visualização, como *Force Plot* e *Decision Plot*, apesar de não utilizados na nossa análise, podem trazer mais informações visuais sobre como os modelos estão tomando decisões. O Force Plot destaca o impacto das variáveis em uma única previsão, enquanto o Decision Plot permite analisar a influência

acumulada das variáveis ao longo da tomada de decisão do modelo, ajudando a entender de forma mais clara o comportamento do modelo em diferentes cenários.

5. CONCLUSÃO

Neste estudo avaliou-se a GPP no Cerrado, por meio de modelos de ML, como foco no RF e ET, tornando possível identificar e interpretar o impacto de variáveis ambientais e climáticas, incluindo incêndios, temperatura da superfície e índice de vegetação, por meio do método SHAP, o qual permitiu maior transparência e confiabilidade nos resultados, mesmo com as limitações de coleta dos dados.

A utilização de modelos de ML e técnica de XAI trouxe um avanço relevante para este estudo, permitindo superar limitações tradicionais associadas a modelos preditivos tidos como *Black Box*. A capacidade de interpretar os impactos de cada variável reforça a aplicabilidade prática dessas técnicas para análises ambientais e planejamento sustentável.

Altos valores para as variáveis de incêndios e temperatura da superfície apresentaram impactos negativos significativos nas previsões, reduzindo a GPP estimada no processo de análise regressiva. Por outro lado, fatores como a fração fotoabsorvível e o índice de vegetação mostraram-se importantes indicadores positivos, sugerindo relevância da vegetação na recuperação e manutenção da GPP. Sobre essa ótica, vale salientar a importância de considerar fatores ambientais e climáticos ao analisar a saúde e a sustentabilidade de ecossistemas complexos.

Este estudo contribui para que futuros trabalhos realizem abordagens incorporando novas variáveis e modelos híbridos, a fim de aprimorar ainda mais a capacidade preditiva e explicativa. A inclusão de variáveis do solo, como umidade, densidade e composição química, isoietas, disponibilidade de nutrientes, poderia fornecer uma visão mais detalhada sobre a relação entre o ambiente físico, os padrões analisados e o possível impacto no GPP. Além disso, a análise de séries temporais permitiria identificar padrões sazonais e tendências, contribuindo para o entendimento de como variáveis ambientais se comportam ao longo do tempo e como esses comportamentos afetam a produtividade e a vulnerabilidade ambiental. Essa abordagem espacial detalhada, combinada com análises temporais, facilitaria a identificação de extremos climáticos, como secas e chuvas intensas, e seus impactos no ecossistema.

Assim, este trabalho não apenas atingiu seus objetivos iniciais, como também abriu caminhos para novas pesquisas e aplicações práticas, consolidando a relevância da ciência de dados para a sustentabilidade e preservação ambiental.

REFERÊNCIAS

ALPAYDIN, Ethem. Uma abordagem geral. In: Introduction to Machine Learning. 4ª ed., MIT Press Bookstore, Cambridge, 2020. Disponível em: http://bvsms.saude.gov.br/bvs/publicacoes/queimadas_incendios_florestais_alerta_risco.pdf. Acesso em: 18 nov. 2024.

ANDERSON, Liana O.; LATORRE, Marcelo L.; SHIMABUKURO, Yosio E. et al. **Sensor MODIS**: Características gerais e aplicações. **Revista Espaço e Geografia**, Brasília (UNB), v. *6*, n. 1, p. 91–121, 2022.

BRASIL. Ministério da Saúde. Secretaria de Vigilância em Saúde. Departamento de Saúde Ambiental, do Trabalhador e Vigilância das Emergências em Saúde Pública. **Queimadas e incêndios florestais**: alerta de risco sanitário e recomendações para a população, Brasília, 2020. 15 p. Disponível em: http://bvsms.saude.gov.br/bvs/publicacoes/queimadas_incendios_florestais_alerta_risco.pdf. Acesso em: 18 nov. 2024.

BREIMAN, Leo. **Random Forests**. Machine Learning, Statistics Department, University of California, Berkeley, Califórnia, v. 45, p. 5–32, 2001. 31 p. Disponível em: ttps://link.springer.com/article/10.1023/A:1010933404324. Acesso em: 18 nov. 2024.

BRITTO, Larissa F. da S.; PACÍFICO, L.; LUDERMIR, Teresa B. **Inferência Automática de Nível de Dificuldade de Receitas Culinárias Utilizando Técnicas de Processamento de Linguagem Natural**. In: XVII Encontro Nacional de Inteligência Artificial e Computacional (ENIAC 2020), Belém, 2020.

COSTA, Marcos de S. **Análise Bibliométrica da Relação entre Fogo e Cerrado**. 41 f. Monografia apresentada à Universidade do Tocantins com fins de obtenção de título de bacharel em Ciências Biológicas, 2023.

DANELICHEN, Victor H.M.; BIUDES, Marcelo S.; VELASQUE, Maísa C.S. et al. Estimating of gross primary production in an Amazon-Cerrado transitional forest using MODIS and Landsat imagery. **Earth Sciences**, **An. Acad. Bras. Ciênc.**, São Paulo, v. 87, n. 3, Set., 2015.

DARPA. Defense Advanced Research Projects Agency. **Explanable Artificial Intelligence (XAI)**. USA, 2016. DARPA. Disponível em: https://onlinelibrary.wiley.com/toc/26895595/2021/2/4. Acesso em: 18 nov. 2024.

DE SANTANA, R; O.; DELGADO, R.C.; SCHIAVETTI, A. The past, present and future of vegetation in the Central Atlantic Forest Corridor, Brazil. **Remote Sensing Applications: Society and Environment**, Chicago, v. 20, p. 100-357, 2020.

DIAS, Fernanda; SUHADOLNIK. Nicolas B.; CAMARGO, Heloísa C. et al. Prevendo o pulso da Amazônia: insights de aprendizado de máquina sobre a dinâmica do desmatamento. **Revista de gestão ambiental**, Rio de Janeiro, v. 362, n. 121359, p. 1-20, jun. 2024.

DOSHI-VELEZ, Finale; KIM, Been. **Towards a Rigorous Science of Interpretable Machine Learning**, 2017. Disponível em: arXiv:1702.08608 [stat.ML]. Acesso em: 18 nov. 2024.

DRUCKER, Harris. **Improving Regressors Using Boosting Techniques**. In: International Conference On Machine Learning (ICML), Nashville. Proceedings. San Francisco: Morgan Kaufmann Publishers, v. 14, p. 107-115, Ago. 1997. Disponível em: https://www.researchgate.net/publication/2424244_Improving_Regressors_Using_Boosting_Technique. Acesso em: 18 nov. 2024.

FLACH, Peter. **Machine Learning**: The Art and Science of Algorithms that Make Sense of Data. Cambridge University Press, Reino Unido, 2012. 409 p.

GALTON, Francis. *Regression Towards Mediocrity in Hereditary Stature*. **Journal of the Anthropological Institute of Great Britain and Ireland**, Grã-Bretanha, v. 15, p. 2460-263, 1886. Disponível em: http://www.stat.ucla.edu/~nchristo/statistics100C/history_regression.pdf. Acesso em: 18 nov. 2024.

GEURTS, Pierre; ERNST, Daniel; WEHENKEL, Louis. Extremely Randomized Trees. **Machine Learning**, Bélgica, v. 63, n. 1, p. 3-42, 2006. Disponível em: https://www.researchgate.net/publication/220343368_Extremely_Randomized_Trees. Acesso em: 20 dez. 2024.

JADAMA, Ansumana F.; TORAY, Modou K. **Ensemble Learning**: Methods, Techniques, Application. University of New Brunswick, 2024. Disponível em: https://www.researchgate.net/publication/381773312_Ensemble_Learning_Methods_Techniques_Application. Acesso em: 23 dez. 2024.

KAMEL, Didan. **Terra Vegetation Indices Monthly L3 Global 1km SIN Grid**. NASA LP DAAC. University of Arizona, Alfredo Huete - University of Technology Sydney and MODAPS SIPS - NASA. MOD13A3 MODIS, 2015. Disponível em: http://doi.org/10.5067/MODIS/MOD13A3.006 (INDICE DE VEGETAÇÃO). Acesso em: 23 dez. 2024.

LI, Luyi; ZENG, Zhenzhong; ZHANG, Guo et al.. Exploring the Individualized Effect of Climatic Drivers on MODIS Net Primary Productivity through an Explainable Machine Learning Framework. **Remote Sens.**, Basileia (CH), v. 14, n. 4401, p. 2-18, p. 2022, 14, 4401. Disponível em: https://doi.org/10.3390/rs1417440. Acesso em: 21 nov. 2024.

LIPOVETSKY, Stan; CONKLIN, Michael W. **Análise de regressão significativa em coeficientes ajustados modelo de valor de Shapley**. 4ª ed., IOS Press, 2010. Disponível em: https://journals.sagepub.com/doi/abs/10.3233/MAS-2010-0170. Acesso em: 21 nov. 2024.

LUDERMIR, Teresa B. Inteligência Artificial e Aprendizado de Máquina: estado atual e tendências. **Estudos Avançados**, v. 35, n. 101, 2021. Disponível em: https://www.scielo.br/j/ea/a/wXBdv8yHBV9xHz8qG5RCgZd/?format=pdf. Acesso em: 20 dez. 2024.

LUNDBERG, Scott; LEE, Su-In.**A Unified Approach to Interpreting Model Predictions**. Nov. 2017. Disponível em: https://arxiv.org/abs/1705.07874. Acesso em: 23 dez. 2024.

MITCHELL, Tom. **Machine Learning**. McGraw-Hill, New York, 1997. 414 p. Disponível em: https://www.cs.cmu.edu/~tom/mlbook.html. Acesso em: 23 dez. 2024.

MONTGOMERY, Douglas C.; PECK, Elizabeth A.; VINING, Geoffrey G. **Introduction to Linear Regression Analysis**, 6ª ed., John Wiley & Sons, New Jersey, 2021. 432 p.

PAIVA, Emerson J.; FERREIRA, João R.; BALESTRASSI, Pedro P. et al. **Erro Quadrático Médio Multivariado Ponderado na Otimização de Múltiplas Respost**as. XXII Encontro Nacional de Engenharia de Produção, Desenvolvimento Sustentável e Responsabilidade Social: As Contribuições da Engenharia de Produção Bento Gonçalves, RS, Brasil, 15 a 18 de outubro de 2012.

PEI, Z., SUN, T., ERIC DAVIDSON, A., WANG, J. The light use efficiency model for estimating gross primary productivity: Enhancements and challenges. **Journal of Plant Ecology**, Reino Unido, v. 1, n. 1, p. 100-112, 2022.

RIBEIRO, Marco. T.; SINGH, Sameer; GUESTRIN, Carlos. **Why Should I Trust You? Explaining the Predictions of Any Classifier**, Anais da Conferência de 2016 do Capítulo Norte-Americano da Associação de Linguística Computacional: Demonstrações, San Diego, California, p. 97-101, 2016.

RUSSELL, Stuart; NORVIG, Peter. **Inteligência Artificial**: Uma Abordagem Moderna. 3ª ed., Rio de Janeiro: Elsevier, 2016. 1.080 p.

SAWYER, Donald; MESQUITA, Beto; COUTINHO, Bruno et al. **Perfil do Ecossistema Hotspot de Biodiversidade do Cerrado**. Critical, Ecosystem Partnership Fund, Rio de Janeiro, 2017. 520 p.

SEBER, George A. F.; LEE, Alan J. **Linear Regression Analysis**. 2ª ed., Wiley online Library, New Jersey, p. 1-16, 2003. Disponível em:

SHAPLEY, Lloyd S. A Value for n-Person Games. In: Contributions to the Theory of Games (AM-28). **Princeton University Press**, New Jersey, v. 2, p. 307-317, 1952. Disponível em: https://www.rand.org/pubs/papers/P295.html. Acesso em: 21 nov. 2024.

SOUSA, Paulo H. de. **Theory of games and economic behavio**r: a ideia de ciência de John von Neumann e Oskar Morgenstern. 2005. 102 f. Dissertação (Mestrado em História da Ciência) - Pontifícia Universidade Católica de São Paulo, São Paulo, 2005. Disponível em: https://sapientia.pucsp.br/handle/handle/1341. Acesso em: 26 dez. 2024.

TIBCO SOFTWARE INC. ***What is Random Forest?*** 2021. Disponível em: https://www.tibco.com. Acesso em: 26 nov. 2025.

TURING, Alan M. *Computing Machinery and Intelligence*. **Mind**, Inglaterra, v. 59, n. 236, p. 433-460, Out. 1950. Disponível em; https://academic.oup.com/mind/article-abstract/LIX/236/433/986238?redirectedFrom=fulltext. Acesso em: 26 nov. 2024.

WAN, Zhengming; HOOK, Simon; HULLEY, Glynn. **MODIS/Terra Land Surface Temperature/Emissivity Daily L3 Global 1km SIN Grid V061 [Conjunto de dados]**. NASA EOSDIS Land Processes Distributed Active Archive Center. 2021. Disponível em: https://doi.org/10.5067/MODIS/MOD11A2.061(TEMPERATURA DA SUPERFICIE TERRESTRE). Acesso em: 26 dez. 2024.

ZHANG, Zhenyu; LI, Xiaoyu; JU, Weimin et. al. Improved estimation of global gross primary productivity during 1981-2020 using the optimized P model. **Sci Total Environ.**, Holanda, v. 838 (pt 2), p. 156-172, 2022.

ZHENG, Y., SHEN, R., WANG, Y. et al. Improved estimate of global gross primary production for reproducing its long-term variation, 1982–2017. **Earth Syst. Sci. Data**, Alemanha, v. 12, p. 2725–2746, 2020.

APÊNDICE 1

VARIÁVEIS	UNIDADE	FONTES	*LINK*
MODIS/TERRA GROSS PRIMARY PRODUCTIVITY 8-DAY L4 GLOBAL 500 M (GPP)	kgC/m²	MOD17A2H v061	https://lpdaac.usgs.gov/products/mod17a2hv006/
MODIS/TERRA LAND SURFACE TEMPERATURE/EMISSIVITY 8-DAY L3 GLOBAL 1 KM	K	MOD11A2 v006	https://lpdaac.usgs.gov/products/mod11a2v006/
MODIS/TERRA LEAF AREA INDEX/FPAR 8-DAY L4 GLOBAL 500 M	%	MOD15A2H v006	https://lpdaac.usgs.gov/products/mod15a2hv006/
MODIS/TERRA VEGETATION INDICES MONTHLY L3 GLOBAL 1 KM	NDVI	MOD13A3 v006	https://lpdaac.usgs.gov/products/mod13a3v006/
MODIS/TERRA LEAF AREA INDEX/FPAR 8-DAY L4 GLOBAL 500 M	m²/m²	MOD15A2H v006	https://lpdaac.usgs.gov/products/mod15a2hv006/
PRECIPITAÇÃO TOTAL HORÁRIO	mm	Estações Meteorológicas INMET	https://portal.inmet.gov.br/dadoshistoricos
PRESSÃO ATMOSFÉRICA AO NÍVEL DA ESTAÇÃO HORÁRIO	mb		https://portal.inmet.gov.br/dadoshistoricos
TEMPERATURA DO AR BULBO SECO HORÁRIA	C		https://portal.inmet.gov.br/dadoshistoricos
UMIDADE RELATIVA DO AR HORÁRIA	%		https://portal.inmet.gov.br/dadoshistoricos
VENTO VELOCIDADE HORÁRIA	m/s		https://portal.inmet.gov.br/dadoshistoricos
RADIAÇÃO GLOBAL	KJ/m²		https://portal.inmet.gov.br/dadoshistoricos
QUANTIDADE DE INCÊNDIOS	Contagem	Projeto Queimadas	https://terrabrasilis.dpi.inpe.br/queimadas/situacao-atual/estatisticas/estatisticas_estados/
ÁREA QUEIMADA CERRADO	Km²	Projeto Queimadas	https://terrabrasilis.dpi.inpe.br/queimadas/aq1km/#nota

Quadro 1. Variáveis, especificações, Fontes e Medidas