Previsão Mensal de Emissão de Dióxido de Carbono na Amazônia Legal

Luísa Ferreira da Silveira (2210875)
Pedro de Almeida Barizon (2211350)
Theo Couto Xavier (2210610)
Vinícius Lucena Bitu Cortez (2210458)

Professora: Paula Medina Maçaira Louro
Monitora: Laura Nascimento Coutinho
Turma: 33A
Data: 06/06/2025

1. Introdução

1.1. Motivação

Considerando o macrotema do projeto **AMazonizAR**, um esforço de inserção da PUC-Rio no contexto da Amazônia [ECOA PUC-RIO 2023], e a premente situação socioambiental que afeta não apenas a Grande Floresta, mas todo o planeta, optou-se pelo seguinte tema para o projeto: ***Na Amazônia Legal, qual a inter-relação entre incêndios florestais, mudanças climáticas e emissão de dióxido de carbono (CO_2)?***

A escolha do tema justifica-se, em primeiro lugar, pela dimensão do impacto climático associado à região amazônica. A Amazônia Legal cobre aproximadamente 5 milhões de km², correspondendo a cerca de 61% do território brasileiro, e se estende por nove estados da federação [IPEA 2008]. Essa vasta região abriga uma das maiores reservas de carbono do planeta. Estima-se que a biomassa florestal da Bacia Amazônica — incluindo biomassa viva, morta e subterrânea — armazene entre 68,8 e 103,2 bilhões de toneladas de carbono [SAATCHI *et al* 2007].

Nesse cenário, o papel regulador da Floresta no clima está ameaçado. De acordo com o Instituto Nacional de Pesquisas Espaciais (INPE), em 2024 foram registrados mais de 140 mil focos de queimadas na Amazônia Legal. Só no estado do Pará, foram mais de 50 mil focos no ano [INPE 2025b]. Queimadas como essas liberam grandes quantidades de CO_2: segundo o relatório do Sistema de Estimativas de Emissões de Gases de Efeito Estufa (SEEG) de 2023, as emissões de mudança no uso da terra e florestas responderam por 48% das emissões brasileiras de gases de efeito estufa em 2022**,** grande parte advinda da Amazônia.

A crise climática também se manifesta localmente. Estudos apontam que a Bacia Amazônica tenderá a aquecer-se mais do que a média global [IPCC 2021], com impactos diretos sobre o regime de chuvas e a frequência de eventos extremos. A seca histórica de 2023, por

exemplo, afetou mais de 500 mil pessoas na região Norte [G1 2023], agravando ainda mais a vulnerabilidade da população local e a suscetibilidade da Floresta a incêndios.

Diante desse contexto crítico, torna-se fundamental compreender, por meio de métodos quantitativos, de que forma fatores climáticos e incêndios florestais se relacionam com a emissão de dióxido de carbono, com o objetivo de fornecer subsídios para ações de monitoramento, prevenção e mitigação. O projeto, ao propor um modelo preditivo para emissões mensais de CO_2 com base em dados ambientais e de queimadas, alinha-se diretamente às diretrizes do AMazonizAR e aos compromissos climáticos nacionais.

1.2. Questão de Pesquisa

A fim de transformar o tema do projeto em algo mais tangível e mensurável, formulou-se esta questão de pesquisa (QP):

A partir dos indicadores ambientais de um estado da Amazônia Legal e seus níveis de incêndios florestais (queimadas) em um dado mês, qual a quantidade de CO_2, em toneladas, emitida por ele na atmosfera nesse mês?

Para o caso de a absorção ser maior que a emissão, convencionou-se uma emissão negativa. Quanto aos estados, serão considerados segundo a classificação do IBGE [2020]:

- Acre (AC);
- Amapá (AP);
- Amazonas (AM);
- Maranhão (MA);
- Mato Grosso (MT);
- Pará (PA);
- Rondônia (RO);
- Roraima (RR);
- Tocantins (TO).

1.3. Cliente a Ser Atendido

Os clientes-alvo deste projeto incluem, em primeiro lugar, os pesquisadores, docentes e estudantes vinculados ao programa AMazonizAR da PUC-Rio, bem como órgãos ambientais e governamentais — como o IBAMA, o ICMBio, secretarias estaduais de meio ambiente e instituições de planejamento e controle climático. A ferramenta desenvolvida poderá também interessar a organizações não governamentais voltadas à conservação e ao clima, e a atores do setor de carbono, como registradoras e certificadoras de créditos.

Esses públicos têm em comum a necessidade de acesso a dados ambientais atualizados e modelagens confiáveis que lhes permitam acompanhar a dinâmica das emissões de CO_2, identificar

padrões críticos, fundamentar decisões estratégicas e comunicar riscos de forma embasada. A demanda por ferramentas que antecipem variações nas emissões é especialmente relevante em um cenário de aceleração da crise climática e crescente cobrança por medidas concretas e mensuráveis.

O modelo preditivo desenvolvido neste projeto tem como principal função estimar mensalmente as emissões de dióxido de carbono (CO_2) em cada estado da Amazônia Legal, com base em dados de queimadas e variáveis climáticas públicas. Essa estimativa permite aos clientes:

- Monitorar de forma antecipada os impactos de eventos climáticos e sazonais sobre as emissões de CO_2, mesmo em períodos em que ainda não haja medição oficial disponível;
- Identificar tendências anômalas ou crescentes de emissão, orientando o disparo de alertas preventivos para equipes de fiscalização, mitigação e resposta;
- Subsidiar políticas públicas ambientais com base em dados empíricos atualizados e espacializados, apoiando decisões de alocação de recursos ou definição de zonas prioritárias;
- Contribuir para o inventário estadual ou regional de emissões, o que pode alimentar relatórios nacionais e fortalecer compromissos internacionais de redução de gases de efeito estufa;
- Apoiar o planejamento e a quantificação de projetos no mercado de créditos de carbono, fornecendo estimativas robustas e periódicas das emissões geradas, o que contribui para avaliações de impacto e para o cálculo de compensações viáveis.

Ao tornar visível e mensurável a relação entre variáveis ambientais e emissões de CO_2, a ferramenta desenvolvida não apenas atende a demandas concretas dos públicos envolvidos, como também produz impacto relevante no ecossistema institucional de resposta à crise climática. Com isso, amplia-se a capacidade técnica, científica e política de atuação sobre o território amazônico, fortalecendo tanto ações locais quanto compromissos globais.

1.4. Objetivos

Os objetivos deste projeto envolvem tanto interesses acadêmicos, centrados na aplicação prática de técnicas contemporâneas de Ciência de Dados, quanto interesses de ordem cidadã, voltados à compreensão e enfrentamento de uma crise ambiental de escala global, com manifestações críticas na região amazônica. A proposta insere-se em um esforço de conciliação entre formação técnica e responsabilidade social, reconhecendo o papel que dados e modelos podem desempenhar na geração de conhecimento e na orientação de decisões públicas.

1.4.1. Objetivo geral

Desenvolver um modelo preditivo capaz de estimar, com base em dados públicos de queimadas e variáveis climáticas, as emissões mensais de dióxido de carbono (CO_2) nos estados da Amazônia Legal, promovendo uma ferramenta útil para monitoramento ambiental e suporte à tomada de decisão.

1.4.2. Objetivos específicos

1. Integrar e tratar bases de dados públicas relacionadas a queimadas e variáveis climáticas, com granularidade mensal e recorte geográfico sobre a Amazônia Legal;

2. Explorar estatisticamente o conjunto de dados unificado, avaliando distribuição, correlações e padrões sazonais relevantes;

3. Construir modelos de regressão supervisionada (inicialmente lineares e depois com abordagens mais flexíveis, se necessário) para prever a emissão de CO_2 a partir dos atributos disponíveis;

4. Comparar diferentes configurações de modelos, utilizando métricas de avaliação como R^2, MAE e MSE, para aferir qualidade preditiva e generalização;

5. Investigar a contribuição relativa de cada variável explicativa, com vistas à identificação de fatores-chave para a emissão de CO_2;

6. Propor, com base nos achados, indicadores compostos que auxiliem no rastreamento e mitigação de emissões, contemplando também fatores climáticos;

7. Documentar o processo e os resultados, visando à reprodutibilidade e à utilidade do projeto por agentes acadêmicos, ambientais e institucionais.

1.4.3. Expectativa de Resultados

O projeto espera gerar, como principal entrega, um modelo preditivo funcional e bem fundamentado, capaz de estimar as emissões mensais de dióxido de carbono (CO_2) nos estados da Amazônia Legal a partir de dados públicos de queimadas e variáveis climáticas.

A partir da análise estatística e da modelagem, é esperada uma correlação positiva forte entre a área queimada e a emissão de CO_2, refletida por coeficiente de correlação superior a 0,70. Ainda que não se antecipe um bom ajuste com modelos simples univariados — como uma regressão linear apenas com a área queimada como preditora —, acredita-se que a inclusão de variáveis ambientais (como temperatura, umidade e velocidade do vento) aumentará

substancialmente o poder explicativo do modelo. Diante disso, o desempenho será considerado satisfatório caso atenda aos seguintes critérios:

- Coeficiente de determinação (R^2) superior a 0,80, indicando boa capacidade de explicação da variância das emissões observadas;

- Erro Absoluto Médio (MAE) inferior ao desvio padrão da variável-alvo, sugerindo previsões mais precisas do que uma simples imputação da média;

- Erro Quadrático Médio (MSE) significativamente inferior à variância da variável-alvo, apontando que os desvios médios ao quadrado são estatisticamente pequenos diante da dispersão natural dos dados.

Além da avaliação quantitativa, espera-se que o modelo possibilite a identificação de variáveis preditoras relevantes, contribuindo para a formulação de indicadores compostos que combinem aspectos de queimadas e clima. Isso poderá apoiar esforços de monitoramento, diagnóstico e resposta ambiental por parte de órgãos públicos e instituições científicas.

Como desdobramento educacional, espera-se que a realização do projeto proporcione aos envolvidos um avanço significativo na maturidade técnica e analítica no domínio da Ciência de Dados. Isso inclui não apenas o uso de ferramentas e algoritmos, mas também o entendimento crítico sobre os dados, a formulação de hipóteses testáveis e a entrega de uma solução prática, reprodutível e relevante do ponto de vista social e ambiental.

1.5. Metodologia a Ser Empregada

Para garantir a reprodutibilidade e a confiabilidade do presente trabalho, foram coletadas bases de dados públicas e abertas, conforme detalhado na seção de ***Obtenção e Tratamento dos Dados***, contendo informações sobre queimadas e variáveis climáticas nos estados da Amazônia Legal com granularidade mensal.

Em seguida, essas informações passaram por um processo de tratamento e enriquecimento, de modo a unificá-las em uma única base de dados consolidada. O tratamento envolveu desde a padronização de formatos até a imputação e a correção de valores faltantes, com atenção à preservação das características temporais e espaciais relevantes para o estudo.

Com o conjunto final obtido, realizou-se uma análise exploratória (descrita na seção de ***Análise Exploratória dos Dados***) para compreender a estrutura do *dataset*, investigar padrões, identificar possíveis *outliers* e avaliar correlações entre as variáveis envolvidas.

A etapa seguinte consistiu na aplicação de dois modelos de aprendizado de máquina supervisionado voltados à tarefa de regressão:

- **Regressão Linear Múltipla**: técnica estatística que modela a relação entre uma variável dependente contínua (neste caso, a emissão mensal de CO_2) e múltiplas variáveis independentes. Assume-se que essa relação é linear, ou seja, pode ser representada por uma combinação linear dos preditores.

- **Árvore de Regressão**: modelo baseado em uma estrutura hierárquica de decisões, em que os dados são divididos em subconjuntos progressivamente mais homogêneos com relação à variável alvo. Tal técnica é não linear e apresenta boa capacidade de capturar interações e padrões complexos entre os atributos.

Para avaliar o desempenho dos modelos de regressão, foram utilizadas métricas amplamente reconhecidas no contexto de tarefas de regressão contínua. Entre elas, destaca-se o Coeficiente de Determinação (R^2), que quantifica a proporção da variância da variável dependente explicada pelo modelo; o Erro Quadrático Médio (MSE), que penaliza com maior intensidade os grandes desvios; e o Erro Absoluto Médio (MAE), que expressa diretamente a magnitude média dos erros em termos da unidade da variável predita. No caso da árvore de regressão, foi aplicada a técnica de validação cruzada com cinco dobras (*cross-validation*), utilizando-se as mesmas métricas em cada partição para gerar uma estimativa mais robusta e generalizável do desempenho. Esse conjunto de métricas permite uma avaliação abrangente e confiável da precisão preditiva dos modelos construídos.

Por fim, os resultados obtidos foram analisados criticamente, com o objetivo de avaliar se os dados disponíveis permitiriam responder de forma satisfatória à questão de pesquisa. Essa avaliação considerou o desempenho dos modelos conforme os critérios estabelecidos na subseção ***Expectativa de Resultados*** — como a capacidade de explicação da variância, a magnitude relativa dos erros e a estabilidade das predições. Também foram examinadas as variáveis com maior relevância preditiva, a fim de verificar em que medida os fatores ambientais, além da área queimada, contribuem de forma significativa para as emissões mensais de CO_2 na Amazônia Legal. Com base nesses elementos, foi possível julgar a adequação dos dados e das técnicas adotadas frente aos objetivos propostos.

2. Metodologia e Plano de Experimentação

2.1. Tipo de Aprendizado de Máquina e Especificação da Tarefa

Dada a natureza da questão de pesquisa — que busca estimar, com base em fatores climáticos e de queimadas, a quantidade mensal de dióxido de carbono (CO_2) emitida na atmosfera por cada estado da Amazônia Legal — optou-se pelo uso de aprendizado de máquina supervisionado, cuja característica principal é o uso de dados rotulados para treinar os modelos.

Mais especificamente, a tarefa em questão é de regressão, uma vez que a variável-alvo, `car_c02_emitido`, é numérica contínua expressa em toneladas. O objetivo central é prever valores reais dessa variável a partir de um conjunto de atributos independentes também quantitativos ou transformados numericamente.

A escolha por um modelo supervisionado de regressão se impôs naturalmente, dado que se teve acesso a séries históricas com emissões já conhecidas para os estados da Amazônia Legal. Assim, havia à disposição pares de entrada e saída — isto é, registros com os fatores ambientais e de queimadas associados aos respectivos valores de emissão de CO_2 —, o que permitiu o treinamento e avaliação de modelos preditivos baseados em dados reais.

Essa abordagem mostrou-se coerente com os objetivos do projeto, que exigem não apenas a compreensão das variáveis envolvidas, mas a capacidade de prever valores futuros ou não observados com base em padrões históricos. Afinal, a regressão é a técnica estatística e computacional adequada quando se deseja estimar uma variável dependente contínua a partir de um conjunto de preditores.

2.2. Técnicas/Algoritmos Utilizados

Para abordar o problema de predição da emissão mensal de CO_2 na Amazônia Legal com base em fatores ambientais e de queimadas, foram aplicadas duas técnicas supervisionadas clássicas: a Regressão Linear Múltipla e a Árvore de Decisão para Regressão. A escolha por essas abordagens foi motivada pela combinação de três critérios principais: adequação técnica ao problema (regressão contínua), interpretabilidade dos modelos e familiaridade do grupo com os algoritmos, o que facilita sua correta implementação e análise crítica.

2.2.1. Regressão Linear Múltipla

A Regressão Linear Múltipla é uma técnica estatística que modela a relação entre uma variável dependente contínua y (neste caso, a emissão de CO_2) e múltiplas variáveis independentes $x_1, x_2, \ldots, x_p$, assumindo uma relação linear entre elas. O modelo é expresso pela equação:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon$$

Nessa equação, β_0 representa o intercepto, β_i os coeficientes de regressão que indicam o impacto de cada variável preditora na variável-alvo, e ε o erro aleatório. A estimação dos coeficientes se dá, usualmente, pelo método dos mínimos quadrados, que minimiza a soma dos quadrados dos resíduos (diferença entre valores reais e previstos).

Essa técnica foi escolhida por sua simplicidade e interpretabilidade, o que facilita a análise dos efeitos individuais das variáveis climáticas e de queimadas sobre as emissões de CO_2. Além disso, tem como aplicação permitir testar a significância estatística de cada variável — dada por seu respectivo coeficiente β_i —, ajudando na seleção de atributos relevantes.

2.2.2. Árvore de Regressão (Decision Tree Regressor)

A Árvore de Regressão é um algoritmo baseado em uma estrutura de decisão hierárquica, em que o espaço dos dados é particionado de forma recursiva com base em divisões que minimizam o erro de predição em cada subgrupo. Cada nó interno da árvore representa uma condição sobre uma variável, enquanto cada folha representa uma predição final (valor médio dos dados que caem naquele nó terminal).

Ao contrário da regressão linear, esse modelo não assume linearidade nem independência entre as variáveis preditoras, o que permite capturar relações não lineares e interações complexas entre os atributos — como, por exemplo, efeitos combinados de temperatura e umidade em diferentes estados.

A árvore de decisão foi selecionada por sua alta interpretabilidade visual e por sua flexibilidade na modelagem de fenômenos que não seguem padrões estritamente lineares, como é comum em processos ambientais e climáticos. Além disso, tem como aplicação permitir a extração da importância relativa das variáveis na formação das decisões, oferecendo *insights* adicionais sobre os fatores mais determinantes para as emissões.

2.2.3. Implementação

Ambos os modelos foram implementados com a biblioteca de Python *scikit-learn*, sendo o de árvore complementado por validação cruzada do tipo *K-Fold* (K = 5) para mitigar riscos de *overfitting* e garantir maior robustez na avaliação.

A adoção simultânea de dois algoritmos com naturezas diferentes — um paramétrico e linear, outro não paramétrico e flexível — permitiu contrastar abordagens e avaliar a capacidade de generalização sob diferentes hipóteses, contribuindo para uma compreensão mais abrangente dos dados e da tarefa de predição proposta.

2.3. Percentual da Base de Dados para Treinamento

Para a avaliação dos modelos de regressão, os dados foram particionados em 70% para treinamento e 30% para teste — proporção amplamente adotada em tarefas supervisionadas por equilibrar a quantidade de dados disponível para o aprendizado com uma amostra suficientemente representativa para validação. Essa divisão foi aplicada tanto à Regressão Linear Múltipla quanto à Árvore de Regressão.

Com o objetivo de assegurar a reprodutibilidade e a comparabilidade dos resultados, foi fixado o mesmo valor de `random_state` (42) na geração dos conjuntos. Isso garantiu que os dados de treino e teste fossem consistentes entre diferentes execuções e entre os dois modelos, viabilizando uma análise comparativa mais robusta.

A Regressão Linear Múltipla foi avaliada com base nessa única divisão (*hold-out*), por se tratar de um modelo paramétrico de baixa variância, menos suscetível a *overfitting* e com comportamento mais estável entre diferentes partições. Já para a Árvore de Regressão, foi adotada validação cruzada do tipo *K-Fold* (K = 5), com o objetivo de mitigar variações causadas por particionamentos específicos e obter uma estimativa mais robusta da performance preditiva do modelo, dada sua natureza não paramétrica e maior flexibilidade.

2.4. Pré-processamento

No pré-processamento dos dados, foram implementados dois *pipelines* distintos, um para cada modelo treinado. Ambos incluíram a conversão da variável categórica `_estado` em variáveis *dummies*, por meio da técnica de *one-hot encoding*, com o parâmetro `drop_first=True`, a fim de evitar multicolinearidade perfeita entre as variáveis geradas. Essa transformação permitiu representar adequadamente as informações categóricas em formato numérico, compatível com os algoritmos utilizados.

No *pipeline* da Regressão Linear Múltipla, foram realizadas etapas adicionais de seleção e padronização de variáveis. Inicialmente, foram removidas variáveis com alta multicolinearidade, com o intuito de evitar instabilidade nos coeficientes estimados e redundância informacional. Para identificar essas relações, utilizou-se a matriz de correlação entre as variáveis preditoras, conforme descrito na subseção ***Análise de Correlação***. Nessa matriz, observaram-se os seguintes coeficientes elevados (ρ):

- `cli_temp_orvalho_med` com `cli_umid_rel_med`: ρ = 0,85
- `cli_umid_rel_min_med` com `cli_umid_rel_min_min`: ρ = 0,83
- `cli_umid_rel_med` com `cli_umid_rel_min_med`: ρ = 0,84
- `cli_umid_rel_med` com `cli_temp_orvalho_med`: ρ = 0,85
- `cli_veloc_vento_med` com `cli_veloc_vento_max`: ρ = 0,78

Com base nesses valores, foram excluídas do modelo as seguintes variáveis explicativas:

- `cli_temp_orvalho_med`
- `cli_umid_rel_med`
- `cli_umid_rel_min_min`
- `cli_veloc_vento_max`

Além disso, foi aplicada a padronização (*z-score normalization*) a todas as variáveis preditoras por meio da classe `StandardScaler` da biblioteca *scikit-learn*. Essa transformação é fundamental para a regressão linear, pois garante que todas as variáveis estejam na mesma escala, o que contribui para maior estabilidade computacional e facilita a interpretação relativa dos coeficientes. A padronização também ajuda a evitar que variáveis com maior magnitude dominem a modelagem, tornando o ajuste numérico mais estável e confiável.

Por fim, vale ressaltar que, embora a matriz de correlação tenha sido utilizada para orientar a exclusão de variáveis altamente colineares, a correlação entre atributos não implica, por si só, uma relação causal. Assim, as exclusões foram guiadas por critérios técnicos de estabilidade do modelo, e não por inferência causal entre os fatores ambientais.

No caso da Árvore de Regressão, não foram aplicadas transformações de escala nem exclusão de variáveis por colinearidade, uma vez que esse modelo não é sensível à magnitude das variáveis nem aos efeitos de multicolinearidade. Assim, o *pipeline* para esse modelo restringiu-se à codificação da variável categórica.

2.5. Métricas de Desempenho

Ambos os modelos foram avaliados com base em métricas amplamente utilizadas para tarefas de regressão contínua, que permitem mensurar o desempenho preditivo de forma precisa e comparável.

As fórmulas a seguir utilizam a seguinte notação:

- n: número total de observações;
- y_i: valor real observado da variável-alvo para a i-ésima amostra;
- $\hat{y}_i$: valor predito pelo modelo para a i-ésima amostra;
- $\bar{y}$: média dos valores reais da variável-alvo.

As métricas adotadas foram:

- **Erro Quadrático Médio (MSE)**: penaliza mais fortemente erros grandes, por elevar ao quadrado a diferença entre o valor real e o valor previsto. É uma métrica sensível a *outliers* e naturalmente minimizada pela regressão linear, cujo critério de otimização é baseado nos mínimos quadrados. Sua fórmula é:

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

- **Erro Absoluto Médio (MAE)**: calcula a média dos módulos dos erros. Por ser uma métrica de ordem 1, é mais robusta a *outliers* e preserva a unidade da variável predita, o que facilita a interpretação direta do erro médio em termos práticos. A fórmula é:

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

- **Coeficiente de Determinação (R^2)**: expressa a proporção da variância da variável dependente explicada pelo modelo. É amplamente adotado por sua interpretabilidade e por ser independente da escala dos dados. Embora muitas vezes esteja entre 0 e 1, o R^2 pode assumir valores negativos, o que indica que o modelo é inferior a uma simples imputação pela média. Sua fórmula é:

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

O uso conjunto dessas métricas proporcionou uma avaliação abrangente e equilibrada do desempenho dos modelos. O MSE foi selecionado por sua coerência com a função de perda minimizada pela regressão linear; o MAE foi empregado por sua maior robustez a valores extremos — comuns em contextos ambientais —; e o R^2, por sua capacidade de mensurar a qualidade do ajuste de forma intuitiva e comparável entre diferentes modelos e *datasets*.

3. Obtenção e Tratamento dos Dados

3.1. Justificativa para a Escolha dos Dados/Atributos

Dado que a variável-alvo do projeto é a emissão de dióxido de carbono (CO_2), sua inclusão no conjunto de dados é essencial. Como a questão de pesquisa define que as previsões devem ter granularidade mensal e estadual, os atributos "ano", "mês" e "estado" também foram incorporados para viabilizar a correta agregação temporal e espacial.

No escopo das queimadas, foram selecionadas duas variáveis principais: a área queimada total (em hectares) e a quantidade de focos de incêndio detectados. A área queimada expressa a extensão do território afetado, enquanto o número de focos indica a frequência dos eventos — permitindo uma descrição conjunta contínua e discreta das queimadas. Ambas as variáveis são reconhecidas como bons indicadores de emissões, pois refletem diretamente o volume de biomassa afetada pelo fogo, fator determinante na liberação de CO_2 e outros gases [Andreae e Merlet 2001].

Outros fatores relacionados foram considerados, como o *Fire Radiative Power* (FRP), que mede a intensidade energética dos incêndios [KAUFMAN *et al* 1996]; o risco de fogo, que estima a probabilidade de ignição com base em condições climáticas; e o número de dias consecutivos sem chuva, variável associada à secura do material combustível [Aragão *et al* 2018]. Contudo, a ausência de séries históricas consolidadas e mensalizadas inviabilizou sua inclusão nesta etapa.

Indicadores socioeconômicos, como taxas de desmatamento e dados fundiários, foram cogitados por sua relação com a ação antrópica nas queimadas [Nepstad *et al* 2006], mas foram deixados de fora por motivos de granularidade incompatível e complexidade adicional de modelagem.

No que se refere aos fatores climáticos, utilizaram-se os dados do Instituto Nacional de Meteorologia [INMET 2025], que disponibiliza registros com alta resolução temporal. Após processamento, foram agregadas as seguintes variáveis mensais: temperatura média, umidade relativa (mínima, máxima e média), pressão atmosférica média e velocidade média e máxima do vento. Tais variáveis estão ligadas à intensidade, à propagação do fogo e à dispersão dos gases emitidos [Seiler e Crutzen 1980], sendo relevantes tanto na explicação das emissões quanto no aumento do poder preditivo dos modelos.

Assim, a seleção final de atributos foi guiada por três critérios: (i) respaldo teórico e empírico na literatura científica; (ii) disponibilidade pública e atualizada dos dados; e (iii) viabilidade técnica de processamento e integração mensal.

3.2. Fontes dos Dados

Foram utilizadas as seguintes fontes, todas públicas e com dados gratuitos para *download*. Uma vez que todas são organizações governamentais ou civis com compromisso socioambiental e científico, acreditou-se serem confiáveis e de qualidade.

3.2.1. Dados de emissões

Banco de Dados do SEEG (Sistema de Estimativas de Emissões de Gases de Efeito Estufa) – O SEEG, mantido pelo Observatório do Clima [2025], fornece estimativas de emissões de gases de efeito estufa, incluindo CO_2, para diferentes setores e regiões do Brasil, inclusive a Amazônia Legal.

3.2.2. Dados de queimadas

MapBiomas – Monitor do Fogo – O Monitor do Fogo, desenvolvido pelo MapBiomas [2025], oferece uma plataforma interativa de monitoramento de áreas queimadas no Brasil, com base em imagens de satélite processadas automaticamente. O sistema permite a visualização e a análise temporal das ocorrências de fogo em diferentes biomas, com alta resolução espacial e atualizações frequentes.

Instituto Nacional de Pesquisas Espaciais (INPE) – Programa Queimadas – O Programa Queimadas do INPE [2025a] fornece dados detalhados sobre focos de calor e áreas queimadas no Brasil e na América Latina, com atualizações quase em tempo real. A plataforma **TerraBrasilis** permite acesso a séries históricas desde 1998, além de mapas, estatísticas e análises espaciais por bioma, estado e município, sendo uma referência nacional no monitoramento de incêndios florestais.

3.2.3. Dados ambientais

Instituto Nacional de Meteorologia (INMET) – O INMET [2025] disponibiliza dados meteorológicos oficiais do Brasil, incluindo medições de temperatura, umidade, pressão atmosférica, velocidade do vento e precipitação. Essas informações, obtidas por meio de uma rede nacional de estações, são fundamentais para análises climáticas e ambientais com granularidade temporal variada, incluindo séries históricas mensais.

Para uma visão mais específica, segue o dicionário de fontes:

Categoria	Espaço	Tempo	Descrição	Fonte
Clima	município	hora	Fatores climáticos (temperatura, precipitação, umidade relativa do ar, dentre outros) a partir de medições horárias de estações meteorológicas em alguns municípios a partir de 2000.	INMET
CO2	município	ano	Emissões de CO2 em toneladas por estado do Brasil por ano a partir de 2000.	Observatório do Clima

Queimadas	estado	mês	Mapeamento mensal de cicatrizes de fogo por estado a partir de 2019.	MapBiomas
Queimadas	estado	mês	Série histórica, mapeada de 1985 a 2023, que apresenta as cicatrizes do fogo mensais e por estado.	MapBiomas
Queimadas	estado	dia, mês, ano	*Fire Radiative Power* (FRP), quantidade de focos de queimadas, risco de fogo e dias sem chuva, todos por estado por mês a partir de 2019.	TerraBrasilis
Queimadas	estado	mês	Quantidade de focos de queimadas por estado por mês a partir de 1998.	TerraBrasilis

Tabela 1 – Dicionário de fontes do projeto

3.3. Procedimentos de Limpeza/Transformação/Redução

Todos os procedimentos descritos nesta seção foram implementados em Jupyter Notebooks no *Google Colab*, utilizando as bibliotecas *pandas* e *numpy*, do ecossistema Python para manipulação e transformação dos dados.

3.3.1. Dados de Emissões

Reformatação do arquivo

O arquivo CSV com os dados do SEEG apresentou formatação incompatível com a leitura direta pelo pandas. Para viabilizar a importação, utilizou-se apoio da ferramenta ChatGPT-4o-mini, com o seguinte prompt: *"Poderia formatar adequadamente o suposto csv que lhe enviarei? Quando fui abri-lo no pandas, os dados não foram processados adequadamente."* A formatação do arquivo foi ajustada com base na resposta recebida, permitindo o carregamento correto dos dados.

Outliers

Quanto a *outliers*, optou-se por deliberadamente não os remover, uma vez que a emissão de dióxido de carbono na atmosfera é sensível à atividade antrópica, de natureza volátil e imprevisível, porque é atrelada a dezenas de variáveis: disposições legais do País, demanda por madeira, governo vigente etc.

Dados faltantes

Como a granularidade temporal dos dados a alimentarem o modelo deveria ser mensal, fez-se, sob sugestão da professora Paula Maçaira, a seguinte extrapolação: não tendo sido possível coletar informações mês a mês sobre as emissões, mas apenas ano a ano; decidiu-se replicar o valor anual ao longo dos meses, a fim de que refletissem o comportamento geral — algo semelhante à imputação da média, mas multiplicada pelo número de meses. Dessa forma, espera-se, o modelo ainda será capaz de identificar padrões, se bem que com menor teor de detalhe. Seja como for, ainda se procuram as informações com a devida granularidade, que poderão ser incorporadas ao projeto futuramente.

3.3.2. Dados de queimadas

Outliers

Novamente, não foram removidos *outliers*, porque poderiam indicar eventos atípicos, que contribuiriam para o aumento das emissões.

Dados faltantes

Quanto à quantidade de focos, não houve dados faltantes. Quanto à área queimada, foram utilizadas duas bases de dados contendo informações por estado e mês: uma principal, abrangendo o período de 1985 a 2023; e outra complementar, com dados disponíveis entre 2019 e 2025. Ambas forneceram registros mensais por unidade federativa, permitindo análises temporais e geográficas das queimadas ao longo de quatro décadas.

Durante o processo de integração, foi identificado que a base principal apresentava valores ausentes nos anos de 2019 a 2023. Para lidar com essas lacunas, optou-se por substituí-las por dados correspondentes da base complementar, considerando a mesma data e estado. Essa decisão foi respaldada por uma análise de correlação entre os dois conjuntos, cujo coeficiente foi de aproximadamente 0,52. Embora não seja elevado, esse valor foi considerado suficiente para justificar o uso de dados observados, ao invés de estimativas com base em médias.

Nos anos anteriores a 2019 (1985 a 2018), em que não havia sobreposição com a base complementar, os valores ausentes foram preenchidos utilizando a média mensal por estado, de modo a preservar padrões sazonais regionais.

Além disso, foram incorporados os registros relativos a 2024 e 2025 provenientes da base complementar, dada a alta correlação desses valores com as médias históricas da base principal (0,88), conferindo robustez à extrapolação temporal do conjunto final.

3.3.3. Dados ambientais

Outliers

Trataram-se os *outliers* após comparar os resultados com e sem o tratamento. Mantendo-os, observaram-se valores incoerentes com a realidade, como temperaturas médias em torno de 12°C, além de um aumento significativo na quantidade de dados faltantes.

Dados faltantes

Em relação aos valores ausentes, optou-se por remover a coluna de radiação, que apresentava mais de 70% dos dados faltantes. Consideramos inviável aplicar técnicas de imputação, tanto pela grande quantidade de valores ausentes quanto pela baixa relevância da radiação para as queimadas na Amazônia. A região apresenta alta umidade ao longo do ano, o que dificulta a ocorrência de queimadas por causas naturais, como o acúmulo de radiação solar.

3.4. Tamanho da Base Final

Após a etapa de tratamento e integração das diferentes fontes de dados, obteve-se um conjunto consolidado com 1.025 observações (linhas) e 15 atributos (colunas). Esse conjunto representa, para cada mês e estado da Amazônia Legal, uma linha com os respectivos valores de emissão de CO_2, variáveis climáticas e indicadores de queimadas.

Nome do Atributo	Descrição	Tipo de Dado
`_ano`	Ano da observação.	Numérico (`Int32`)
`_estado`	Unidade federativa.	Categórico (`str`)
`_mes`	Mês da observação.	Numérico (`Int32`)
`car_co2_emitido`	Emissão de CO_2 (toneladas).	Numérico (`Float64`)
`cli_pressao_atm_med`	Pressão atmosférica media.	Numérico (`Float64`)
`cli_temp_ar_med`	Temperatura média do ar.	Numérico (`Float64`)
`cli_temp_orvalho_med`	Temperatura média no ponto de orvalho.	Numérico (`Float64`)
`cli_umid_rel_med`	Umidade relativa media.	Numérico (`Float64`)
`cli_umid_rel_min_max`	Umidade relativa mínima (valor máximo).	Numérico (`Float64`)
`cli_umid_rel_min_med`	Umidade relativa mínima (valor médio).	Numérico (`Float64`)
`cli_veloc_vento_max`	Velocidade máxima do vento.	Numérico (`Float64`)
`cli_veloc_vento_med`	Velocidade média do vento.	Numérico (`Float64`)
`que_area_queimada`	Área queimada total (hectares).	Numérico (`Float64`)
`que_focos_qtd`	Quantidade de focos de queimadas.	Numérico (`Float64`)

Tabela 2 – Atributos do *dataset* resultante do tratamento

As Figuras 1 e 2 abaixo ilustram, respectivamente, a primeira e a segunda metade das colunas da base final. Note que o *shape* completo da base — (1025, 15) — pode ser visualizado no canto inferior direito da Figura 1.

	_ano	_estado	_mes	car_co2_emitido	cli_pressao_atm_med	cli_temp_ar_med	cli_temp_orvalho_med	cli_umid_rel_med	cli_umid_rel_min_max	cli_umid_rel_min_med
0	2008	AC	7	2.627698e+07	986.843612	28.142731	18.914978	59.555066	95.0	54.432558
1	2008	AC	9	2.627698e+07	991.705941	24.446194	19.467987	75.811881	97.0	72.496700
2	2008	AC	10	2.627698e+07	990.328360	25.229298	21.617473	81.870968	96.0	78.720430
3	2008	AC	11	2.627698e+07	988.610987	25.195410	22.624478	86.905424	96.0	84.048679
4	2008	AC	12	2.627698e+07	988.692608	24.898790	22.727554	88.529570	96.0	86.116935

Shape: (1025, 15)

Figura 1 – Primeira metade do *dataset* após o tratamento

cli_umid_rel_med	cli_umid_rel_min_max	cli_umid_rel_min_med	cli_umid_rel_min_min	cli_veloc_vento_max	cli_veloc_vento_med	que_area_queimada	que_focos_qtd
59.555066	95.0	54.432558	29.0	5.1	2.152915	4957.0	165.0
75.811881	97.0	72.496700	25.0	1.0	0.210504	46073.0	2947.0
81.870968	96.0	78.720430	29.0	1.0	0.204959	30355.0	856.0
86.905424	96.0	84.048679	42.0	1.0	0.186970	2082.0	63.0
88.529570	96.0	86.116935	53.0	1.0	0.179442	127.0	4.0

Figura 2 – Segunda metade do *dataset* após o tratamento

3.5. Dicionário de Dados

A seguir, apresenta-se o dicionário de dados do projeto, que contém todas as variáveis utilizadas após o tratamento. Para cada atributo, indicam-se o nome, o tipo, a descrição, a unidade de medida e a fonte de que foi extraído. Os atributos que compõem a chave primária do *dataset* foram grafados em vermelho.

Nome	Tipo	Descrição	Unidade	Fonte
_ano	Int32	Ano segundo o calendário cristão.	ano	Não se aplica
_estado	string	Nome do estado da Amazônia Legal.	adimensional	Não se aplica
_mes	Int32	Mês do ano segundo o calendário cristão. Janeiro relaciona-se com 1, fevereiro com 2 e assim em diante.	mês	Não se aplica
car_c02_emitido	Float64	Quantidade de dióxido de carbono emitido na atmosfera. Se for negativa, considera-se que tenha havido absorção.	toneladas (t)	SEEG
cli_pressao_atm_med	Float64	Média das medições mensais de pressão atmosférica de todas as estações meteorológicas do estado disponíveis no Instituto Nacional de Meteorologia (INMET).	mb (milibar)	INMET
cli_temp_ar_med	Float64	Média das medições mensais de temperatura do ar de todas as estações meteorológicas do estado disponíveis no Instituto Nacional de Meteorologia (INMET).	°C	INMET
cli_temp_orvalho_med	Float64	Média das medições mensais de temperatura do ponto de orvalho de todas as estações meteorológicas do estado disponíveis no Instituto Nacional de Meteorologia (INMET).	°C	INMET
cli_umid_rel_med	Float64	Média das medições mensais de umidade relativa do ar de todas as estações meteorológicas do estado disponíveis no Instituto Nacional de Meteorologia (INMET).	percentual (%)	INMET
cli_umid_rel_min_max	Float64	Máxima dentre as mínimas mensais de umidade relativa do ar de todas as estações meteorológicas do estado disponíveis no Instituto Nacional de Meteorologia (INMET).	percentual (%)	INMET
cli_umid_rel_min_med	Float64	Média das mínimas mensais de umidade relativa do ar de todas as estações meteorológicas do estado disponíveis no Instituto Nacional de Meteorologia (INMET).	percentual (%)	INMET
cli_umid_rel_min_min	Float64	Mínima dentre as mínimas mensais de umidade relativa do ar de todas as estações meteorológicas do estado disponíveis no Instituto Nacional de Meteorologia (INMET).	percentual (%)	INMET

cli_veloc_vento_max	Float64	Máxima dentre as medidas mensais de velocidade do ar de todas as estações meteorológicas do estado disponíveis no Instituto Nacional de Meteorologia (INMET).	m/s	INMET
cli_veloc_vento_med	Float64	Média das medidas mensais de velocidade do ar de todas as estações meteorológicas do estado disponíveis no Instituto Nacional de Meteorologia (INMET).	m/s	INMET
que_area_queimada	Float64	Número de hectares de área queimada.	hectare (Ha)	MapBiomas
que_focos_qtd	Float64	Quantidade de focos de queimada. Se bem que, em essência, seja um valor inteiro, foi considerado um domínio de ponto flutuante por conta de imputação de dados.	adimensional	TerraBrasilis

Tabela 3 – Dicionário de dados do projeto

4. Análise Exploratória dos Dados

Para sistematizar a análise, adotou-se, a nível de código, o seguinte critério: *Se uma dada variável apresentar tipo numérico (`int`, `float`, dentre outros), automaticamente será quantitativa. A cardinalidade será, então, dada pela do tipo: inteiros são contáveis (discretos), e pontos flutuantes representam intervalos reais, não contáveis (contínuos). Caso contrário, assume-se ser qualitativa, e a subcategoria é avaliada ad hoc por análise humana.*

Diante disso, vale ressaltar que, embora `_mes` fosse originalmente uma variável qualitativa ordinal, com domínio (janeiro, fevereiro, ...), com a codificação real para (1, 2, ...), pode-se interpretá-la como a quantidade de meses desde o início do ano, algo quantitativo discreto. Quanto a `_estado`, como não há ordem explícita, segue que é quantitativa nominal. Assim, as variáveis terão a seguinte classificação:

Variável	Tipo de Dado
`_ano`	Quantitativo discreto
`_estado`	Qualitativo nominal
`_mes`	Quantitativo discreto
`car_c02_emitido`	Quantitativo contínuo
`cli_pressao_atm_med`	Quantitativo contínuo
`cli_temp_ar_med`	Quantitativo contínuo
`cli_temp_orvalho_med`	Quantitativo contínuo
`cli_umid_rel_med`	Quantitativo contínuo
`cli_umid_rel_min_max`	Quantitativo contínuo
`cli_umid_rel_min_med`	Quantitativo contínuo
`cli_umid_rel_min_min`	Quantitativo contínuo
`cli_veloc_vento_max`	Quantitativo contínuo
`cli_veloc_vento_med`	Quantitativo contínuo
`que_area_queimada`	Quantitativo contínuo
`que_focos_qtd`	Quantitativo contínuo

Tabela 4 – Classificação das variáveis

4.1. Estatísticas Descritivas

A análise descritiva visa entender o comportamento das variáveis quantitativas mais relevantes para o problema, como emissão de CO_2, focos e área de queimadas, além de variáveis ambientais. Foram utilizadas medidas de tendência central, dispersão, posição e forma da distribuição para detectar padrões, *outliers* e características que podem impactar a modelagem.

4.1.1. Tendência Central

Foram calculadas as seguintes métricas para representar o comportamento central dos dados:

- **Média:** Valor médio da distribuição.
- **Mediana:** Valor que divide a amostra em duas partes iguais (50% dos dados).
- **Moda:** Valor mais frequente (ou múltiplas modas, se aplicável).

Análises

A `car_c02_emitido` apresenta média (`1.075e+08`) significativamente maior que a mediana (`5.667e+07`), sugerindo assimetria à direita (valores extremos elevados). Para `que_focos_qtd`, a mediana (`26.0`) é muito inferior à média (`571.35`), indicando alta concentração de valores baixos com *outliers* positivos.

4.1.2. Dispersão

As métricas de variabilidade incluem:

- **Amplitude:** Diferença entre máximo e mínimo.
- **Variância e Desvio Padrão:** Medem a dispersão em torno da média.
- **Coeficiente de Variação (CV):** Razão entre desvio padrão e média (útil para comparar variáveis em escalas distintas).

Análises

`que_area_queimada` e `que_focos_qtd` apresentam CV > `2.0`, indicando alta dispersão relativa (dados heterogêneos). `_ano` tem CV ≈ `0.0026`, confirmando baixa variabilidade (dados concentrados em um período específico).

4.1.3. Posição (Quartis)

Os quartis (Q1, Q2 = mediana, Q3) dividem os dados em intervalos de 25%. Exemplo: para `cli_temp_ar_med`, 25% dos valores estão abaixo de `25.46 °C` (Q1), enquanto 75% estão abaixo de `27.30 °C` (Q3).

Análises

Em `cli_veloc_vento_med`, a diferença entre Q3 (`1.479`) e Q1 (`0.824`) sugere dispersão moderada na velocidade média do vento.

4.1.4. Forma da Distribuição

Assimetria (Coeficiente de Fisher-Pearson)

Valores próximos de zero indicam simetria, enquanto:

- Assimetria positiva (cauda à direita):
  - `car_c02_emitido` (`2.106`), `que_focos_qtd` (`2.802`), e `que_area_queimada` (`5.136`) mostram *outliers* elevados.

- Assimetria negativa (cauda à esquerda):
  - `cli_pressao_atm_med` (`-2.161`) e `cli_temp_orvalho_med` (`-2.164`) têm concentração de valores altos.

Curtose (Coeficiente de Fisher)
Mede o "achatamento" da distribuição:

- Leptocúrtica (curtose > 0): Caudas pesadas e pico agudo.
  - `cli_umid_rel_min_max` (`45.515`) e `que_area_queimada` (`39.861`) indicam *outliers* extremos.

- Platicúrtica (curtose < 0): Distribuição mais achatada (ex.: `_ano`, `_mes`).

Observação: Essas métricas foram calculadas usando `Series.skew()` e `Series.kurtosis()` do pandas.

4.1.5. Conclusão

As estatísticas descritivas revelam padrões importantes, como as variáveis ambientais (ex.: `cli_temp_ar_med`) tenderem a ser simétricas (assimetria próxima de zero) e os dados de emissão de CO_2 e queimadas apresentarem assimetria positiva e curtose elevada, confirmando a presença de *outliers*. Isso sugere a necessidade de transformações ou métodos robustos para modelagem. Para maiores detalhes, seguem as medidas estatística abaixo.

Métricas: CENTRALITY

	media	mediana	moda
_ano	2014.468293	2014.0	[2015, 2019]
_mes	6.500488	7.0	[1, 5]
car_c02_emitido	107520273.652189	56669772.946213	[3652935.022779369, 3705752.691345386, 3958983...
cli_pressao_atm_med	993.395296	997.229524	[1007.8279069767442, 1011.1125]
cli_temp_ar_med	26.429096	26.29328	[21.97972222222222, 21.99903181189488, 22.0466...
cli_temp_orvalho_med	21.282504	22.111877	[20.767185289957567, 22.77238493723849]
cli_umid_rel_med	75.905264	79.0	83.438172
cli_umid_rel_min_max	94.459512	95.0	95.0
cli_umid_rel_min_med	72.934699	75.870968	[71.88888888888889, 81.89516129032258]
cli_umid_rel_min_min	38.312195	39.0	34.0
cli_veloc_vento_max	3.862341	3.6	3.4
cli_veloc_vento_med	1.245765	1.155405	[0.0163475699558173, 0.0483660130718954, 0.055...
que_area_queimada	109144.90887	14356.0	[958.0, 12673.4]
que_focos_qtd	571.351247	26.0	1.0

Figura 3 – Medidas de centralidade das variáveis quantitativas

Métricas: DISPERSION

	max	min	amplitude	variancia	desv_pad	coef_var
_ano	2.023000e+03	2.000000e+03	2.300000e+01	2.796799e+01	5.288477e+00	0.002625
_mes	1.200000e+01	1.000000e+00	1.100000e+01	1.192798e+01	3.453691e+00	0.531297
car_c02_emitido	6.532973e+08	3.271708e+06	6.500256e+08	1.794343e+16	1.339531e+08	1.245840
cli_pressao_atm_med	1.012780e+03	9.229797e+02	8.980061e+01	2.922557e+02	1.709549e+01	0.017209
cli_temp_ar_med	3.176319e+01	2.197972e+01	9.783472e+00	2.106806e+00	1.451484e+00	0.054920
cli_temp_orvalho_med	2.598745e+01	7.230854e+00	1.875659e+01	7.506876e+00	2.739868e+00	0.128738
cli_umid_rel_med	9.239058e+01	1.696992e+01	7.542065e+01	1.358411e+02	1.165509e+01	0.153548
cli_umid_rel_min_max	1.000000e+02	3.000000e+01	7.000000e+01	2.631306e+01	5.129625e+00	0.054305
cli_umid_rel_min_med	9.186012e+01	1.496162e+01	7.689850e+01	1.421618e+02	1.192316e+01	0.163477
cli_umid_rel_min_min	7.700000e+01	9.000000e+00	6.800000e+01	1.612091e+02	1.269681e+01	0.331404
cli_veloc_vento_max	9.600000e+00	6.000000e-01	9.000000e+00	1.796608e+00	1.340376e+00	0.347037
cli_veloc_vento_med	4.102525e+00	1.634757e-02	4.086178e+00	4.466144e-01	6.682922e-01	0.536451
que_area_queimada	3.144010e+06	1.000000e+00	3.144009e+06	6.311846e+10	2.512339e+05	2.301838
que_focos_qtd	7.669000e+03	1.000000e+00	7.668000e+03	1.448882e+06	1.203695e+03	2.106751

Figura 4 – Medidas de dispersão das variáveis quantitativas

Métricas: POSITION

	q1	q2	q3
_ano	2.010000e+03	2.014000e+03	2.019000e+03
_mes	4.000000e+00	7.000000e+00	9.000000e+00
car_c02_emitido	3.130998e+07	5.666977e+07	1.148093e+08
cli_pressao_atm_med	9.861975e+02	9.972295e+02	1.004590e+03
cli_temp_ar_med	2.546319e+01	2.629328e+01	2.730413e+01
cli_temp_orvalho_med	2.073703e+01	2.211188e+01	2.293194e+01
cli_umid_rel_med	7.040599e+01	7.900000e+01	8.394602e+01
cli_umid_rel_min_max	9.400000e+01	9.500000e+01	9.700000e+01
cli_umid_rel_min_med	6.730556e+01	7.587097e+01	8.109489e+01
cli_umid_rel_min_min	3.000000e+01	3.900000e+01	4.700000e+01
cli_veloc_vento_max	3.000000e+00	3.600000e+00	4.700000e+00
cli_veloc_vento_med	8.243733e-01	1.155405e+00	1.479324e+00
que_area_queimada	2.954000e+03	1.435600e+04	8.966600e+04
que_focos_qtd	3.500000e+00	2.600000e+01	3.660000e+02

Figura 5 – Medidas de posição das variáveis quantitativas

Métricas SKEW:

	assimetria
_ano	-0.200539
_mes	-0.004483
car_c02_emitido	2.105874
cli_pressao_atm_med	-2.160808
cli_temp_ar_med	0.369335
cli_temp_orvalho_med	-2.164032
cli_umid_rel_med	-1.366295
cli_umid_rel_min_max	-5.131046
cli_umid_rel_min_med	-1.274984
cli_umid_rel_min_min	-0.096313
cli_veloc_vento_max	0.679577
cli_veloc_vento_med	1.374271
que_area_queimada	5.136334
que_focos_qtd	2.802279

Figura 6 – Medida de assimetria das variáveis quantitativas

Métricas KURTOSIS:

	curtose
_ano	-0.680608
_mes	-1.214559
car_c02_emitido	3.742102
cli_pressao_atm_med	5.872654
cli_temp_ar_med	0.565680
cli_temp_orvalho_med	5.283790
cli_umid_rel_med	2.001775
cli_umid_rel_min_max	45.514632
cli_umid_rel_min_med	1.744670
cli_umid_rel_min_min	-0.050843
cli_veloc_vento_max	1.166441
cli_veloc_vento_med	2.847532
que_area_queimada	39.861126
que_focos_qtd	8.632749

Figura 7 – Medida de curtose das variáveis quantitativas

4.2. Visualizações de Distribuição

O gráfico KDE revela a forma da distribuição das emissões de CO_2: observa-se que a maior densidade de probabilidade ocorre para valores inferiores a 200 milhões de toneladas por mês, o que está alinhado com a média (107,5 milhões) e a mediana (56,7 milhões) calculadas anteriormente. A distribuição apresenta assimetria positiva, evidenciada pela cauda alongada à direita, que confirma a presença de valores extremamente altos (acima de 600 milhões), embora com baixa probabilidade. Além disso, o pico estreito e as caudas relativamente pesadas são compatíveis com o coeficiente de curtose positivo (3,74), caracterizando uma distribuição leptocúrtica, ou seja, com maior concentração de dados próximos à média em comparação com uma distribuição normal. Observa-se também que a densidade é próxima de zero para emissões abaixo de zero, o que descarta a possibilidade de absorção líquida de CO_2 nos dados analisados. Por fim, vale destacar que o KDE suaviza os dados, o que pode mascarar variações locais importantes que seriam visíveis em um histograma.

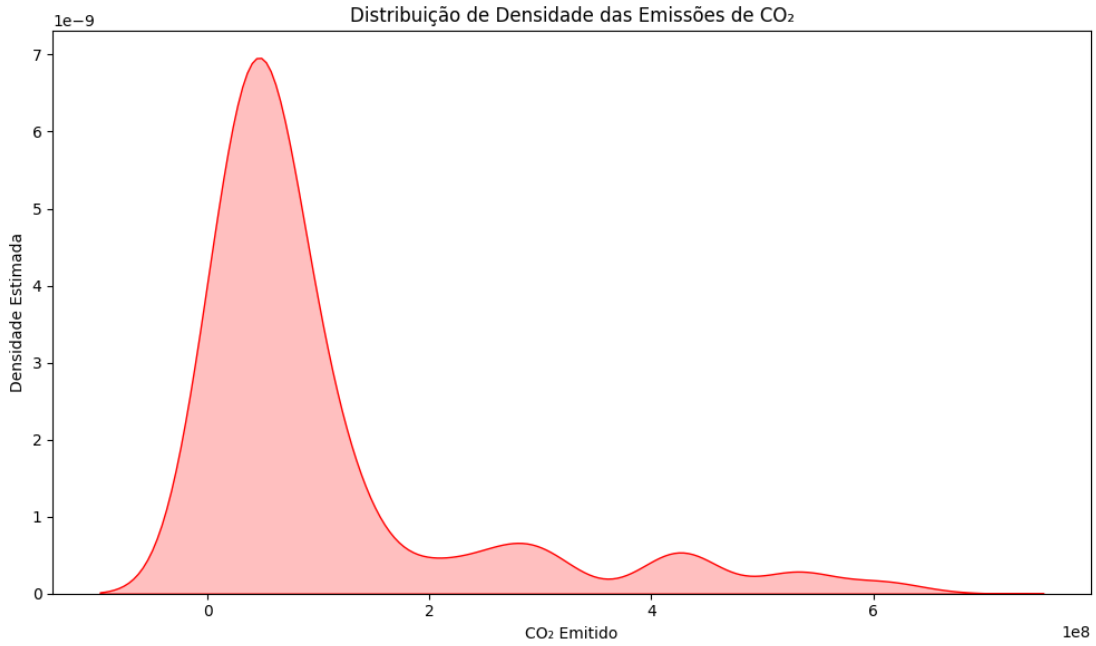

Figura 8 – Gráfico de densidade (KDE) da emissão de CO_2

O histograma complementa o gráfico de KDE ao mostrar a distribuição real das emissões de CO_2 em intervalos discretos. A maior parte das observações, correspondendo a mais de 80% dos dados segundo a contagem de *bins*, está concentrada abaixo de 200 milhões de toneladas, enquanto pouquíssimos valores ultrapassam os 600 milhões, apenas um ou dois *bins* apresentam frequências próximas de zero nessa faixa. Visualmente, observa-se uma assimetria positiva, com a barra mais alta localizada à esquerda e uma cauda longa e esparsa à direita, o que reforça o coeficiente de assimetria de 2,106. Não há registros em *bins* abaixo de zero, o que confirma a ausência de valores negativos e valida a conclusão já apontada pelo KDE. A forma do histograma

também se alinha às métricas estatísticas discutidas na Seção 4.1, especialmente no que diz respeito à assimetria e à curtose. Além disso, a amplitude de 650 milhões e o desvio padrão elevado ($1{,}34 \times 10^8$) ajudam a explicar a dispersão observada nas extremidades da distribuição.

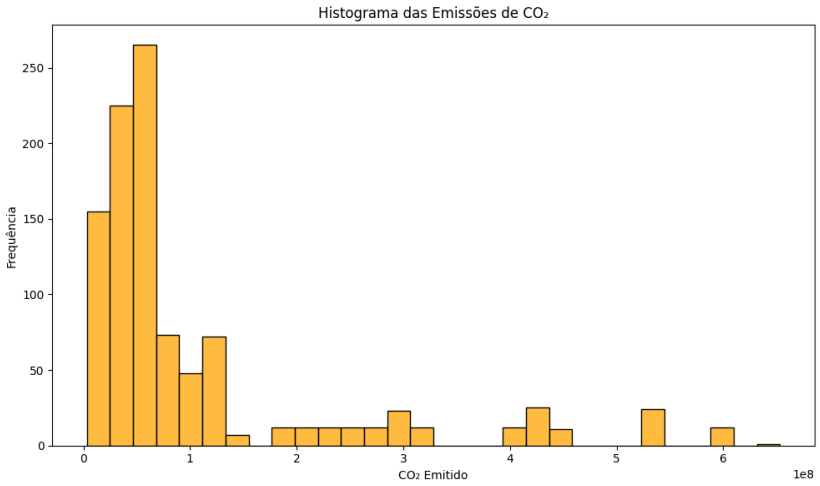

Figura 9 – Histograma da emissão de CO_2

4.3. Análise de Correlação

Consideram-se como pertinentes as correlações com valores absolutos maiores ou iguais a 0.70, segundo o critério usual para indicar correlação forte. Correlações fracas ou próximas de zero (entre -0.05 e 0.05) também são destacadas por sua irrelevância estatística, assim como aquelas próximas do corte de 0.70, que podem indicar relações latentes interessantes.

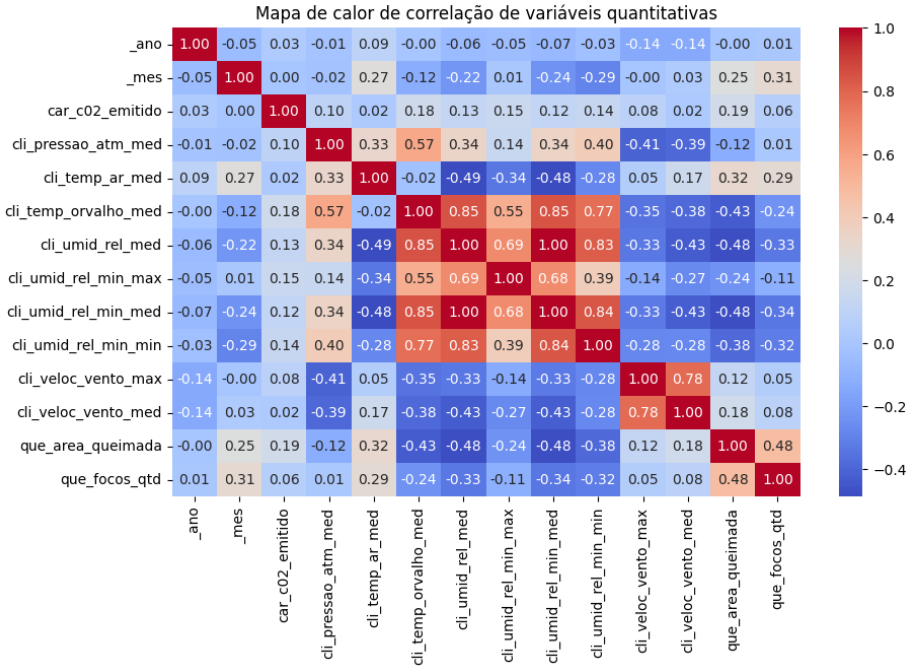

Figura 10 – Mapa de calor de correlação de variáveis quantitativas

4.3.1. Correlações Fracas (|ρ| ≤ 0.05)

As seguintes variáveis apresentam correlação insignificante, indicando ausência de relação linear:

`_ano` ↔ **`cli_pressao_atm_med`** (ρ = -0.01): A pressão atmosférica não variou sistematicamente ao longo dos anos.

`_mes` ↔ **`car_c02_emitido`** (ρ = 0.00): Não há evidência de sazonalidade nas emissões de CO_2.

`_ano` ↔ **`cli_temp_orvalho_med`** (ρ = -0.00): A temperatura de orvalho manteve-se estável temporalmente.

`cli_veloc_vento_med` ↔ **`car_c02_emitido`** (ρ = 0.02): A velocidade média do vento não influencia as emissões.

Observação: Correlações próximas de zero não descartam relações não-lineares ou contextos específicos não capturados.

4.3.2. Correlações Moderadas (0.05 < |ρ| < 0.70)

Relações parciais que exigem contextualização:

`que_area_queimada` ↔ **`car_c02_emitido`** (ρ = 0.19): A baixa correlação contraria a expectativa teórica, possivelmente devido à dispersão temporal (dados anuais imputados para meses). Quando agregados por estado, a correlação sobe para 0.57, revelando distorções na escala mensal.

`que_area_queimada` ↔ **`que_focos_qtd`** (ρ = 0.48): Indica que mais focos tendem a aumentar a área queimada, mas a variabilidade sugere outros fatores em jogo (ex.: tipo de vegetação, ações de combate).

`cli_umid_rel_med` ↔ **`cli_umid_rel_min_max`** (ρ = 0.69): Reflete a coerência interna dos dados climáticos: meses úmidos têm mínimas diárias menos extremas.

4.3.3. Correlações Fortes (|ρ| ≥ 0.70)

Variáveis climáticas:

`cli_veloc_vento_max` ↔ **`cli_veloc_vento_med`** (ρ = 0.78): Velocidades máximas e médias do vento são diretamente proporcionais.

`cli_umid_rel_med` ↔ **`cli_temp_orvalho_med`** (ρ = 0.85): A temperatura de orvalho é função da umidade, como esperado na termodinâmica.

Umidade relativa:
Correlações acima de 0.77 entre médias e mínimas (ex.: `cli_umid_rel_med` ↔ `cli_umid_rel_min_min`) confirmam a estabilidade espacial da umidade.

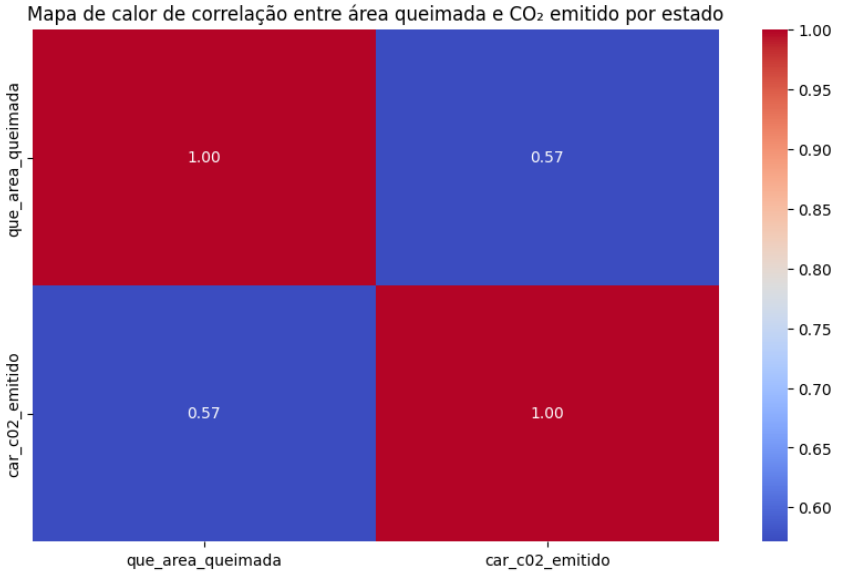

Figura 11 – Correlação entre área queimada e CO_2 emitido (por estado)

4.4. Gráficos de Dispersão

Este gráfico de dispersão mostra a relação entre a quantidade de focos de queimada e a emissão de CO_2 nos diferentes estados da Amazônia Legal. De forma geral, não se observa uma correlação linear forte e direta entre essas duas variáveis. Embora haja uma concentração de pontos com baixos valores tanto para focos quanto para emissão de CO_2, percebe-se uma grande dispersão nos dados.

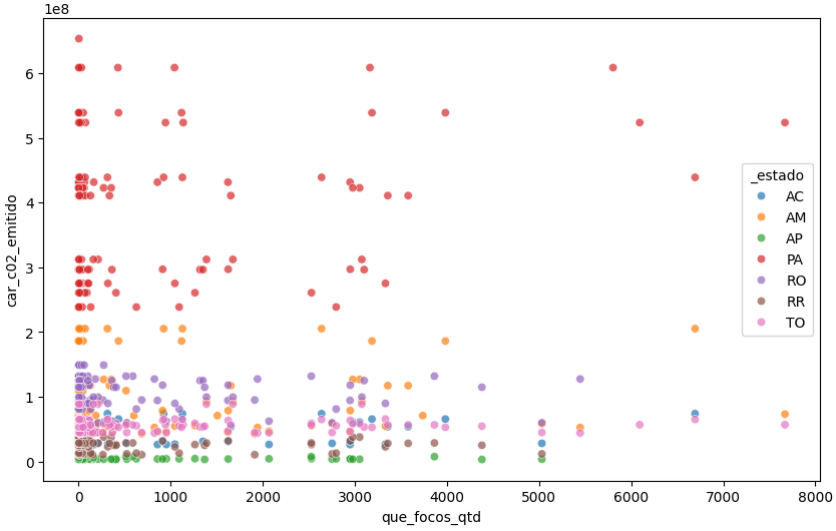

Figura 12 – Focos de queimada *vs* CO_2 emitido

O gráfico de dispersão revela uma relação não linear entre o número de focos de queimadas e as emissões de CO_2 nos estados da Amazônia Legal, evidenciando padrões distintos entre os estados. O Pará, por exemplo, apresenta pontos com poucos focos, mas emissões elevadas, o que pode indicar queimadas de alta intensidade, presença de grande biomassa queimada por foco ou até outras fontes de emissão não diretamente associadas aos focos detectados por satélites. Já o Amazonas e o Acre mostram emissões mais baixas, mesmo com alguma variação no número de focos, possivelmente devido a menor biomassa disponível, queimadas menos intensas ou maior umidade da vegetação, que reduz a eficiência da queima. Esses padrões sugerem que a quantidade de focos, por si só, não explica adequadamente as emissões observadas. Fatores contextuais, como o tipo de vegetação, umidade do solo, estágio do desmatamento e até limitações na detecção por satélite — como a cobertura de nuvens — podem exercer influência significativa sobre os dados. Assim, a relação entre focos de queimadas e emissões de CO_2 se mostra complexa e mediada por múltiplas variáveis ambientais e metodológicas, o que reforça a necessidade de análises complementares com dados sobre biomassa, cobertura vegetal e condições climáticas locais.

Este outro gráfico revela a relação entre a quantidade de focos de queimada e a área queimada nos estados da Amazônia Legal. De maneira geral, observa-se uma correlação positiva, indicando que um maior número de focos tende a estar associado a uma maior área queimada, embora essa relação apresente uma considerável variabilidade.

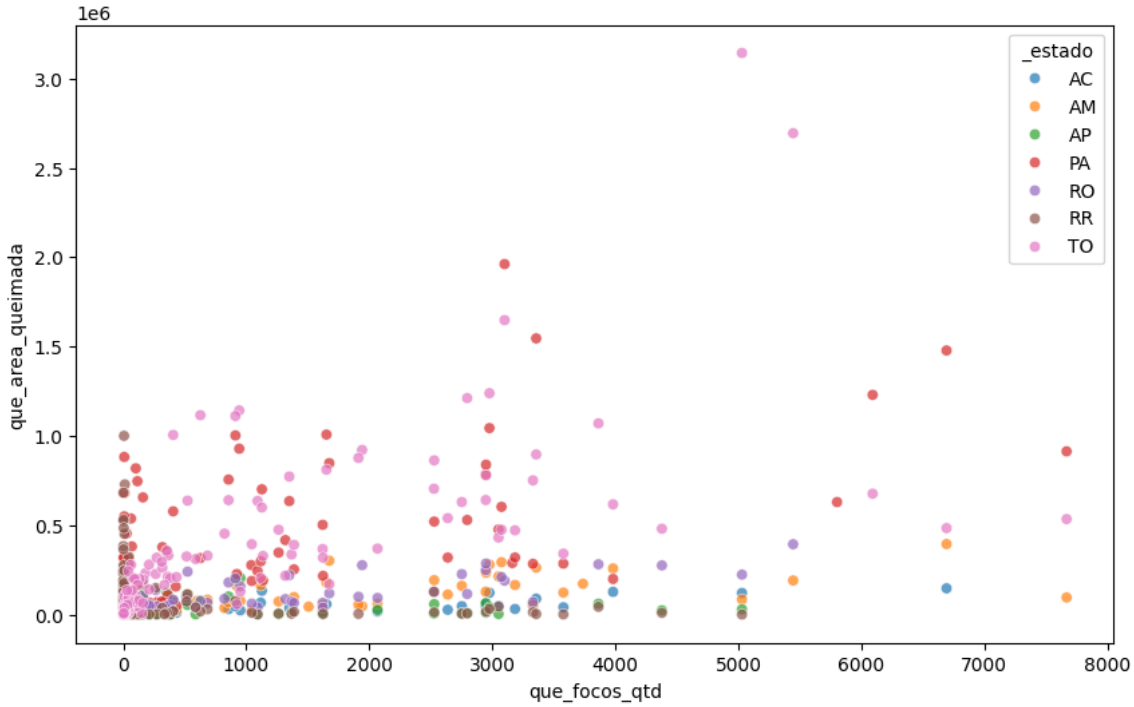

Figura 13 – Área queimada *vs* focos de queimada

O segundo gráfico revela uma correlação positiva entre o número de focos de queimadas e a área queimada, mas essa relação apresenta variações significativas entre os estados da Amazônia Legal. Observa-se uma maior concentração de pontos na parte inferior esquerda do gráfico, indicando que a maioria dos eventos envolve poucos focos e resulta em áreas queimadas relativamente pequenas. No entanto, ao se analisar por estado, padrões distintos emergem. O Pará e o Tocantins, por exemplo, apresentam ocorrências extremas, com muitos focos e áreas queimadas extensas, o que pode refletir dinâmicas associadas à fronteira agrícola, à presença de vegetação mais inflamável ou à incidência de secas severas. Por outro lado, estados como o Acre e o Amapá concentram-se em eventos de menor magnitude, possivelmente em função de políticas ambientais mais restritivas, menor pressão de desmatamento ou características geográficas específicas. A dispersão dos dados, portanto, evidencia que o número de focos, embora relacionado à área queimada, não é um indicador suficiente por si só. Fatores como o tipo de cobertura vegetal, a duração dos incêndios e condições climáticas locais também influenciam significativamente os resultados. Isso sugere a necessidade de análises mais detalhadas, que incorporem variáveis de controle — como bioma, precipitação ou uso do solo — e investiguem casos atípicos, como áreas com grandes queimadas e poucos focos detectados, que podem indicar subnotificação ou falhas nos métodos de medição.

4.5. Análise de Dados Categóricos

Como visto ao início da seção, a única variável categórica é `_estado`, cuja análise foi feita por meio da seguinte tabela de frequência:

Estado	Frequência Absoluta	Frequência Relativa (%)	Freq. Absoluta Acumulada	Freq. Relativa Acumulada (%)
AC	113	11,02	113	11,02
AM	183	17,85	296	28,88
AP	114	11,12	410	40,00
MA	0	0,00	410	40,00
MT	0	0,00	410	40,00
PA	156	15,22	566	55,22
RO	134	13,07	700	68,29
RR	121	11,80	821	80,10
TO	204	19,90	1025	100,00

Tabela 5 – Tabela de frequência de `_estado`

Verifica-se que Amazonas e Tocantins são os estados mais representados neste dataset, enquanto Maranhão e Mato Grosso sequer foram representados, cuja ausência explica-se pela dificuldade em coletar suas informações climáticas, o que impossibilitou sua aparição no *dataset* final. Seja como for, uma vez que a quantidade de dióxido de carbono emitida na atmosfera não deveria depender diretamente da posição geográfica, mas sim da área efetivamente queimada, provavelmente não serão necessárias nem a sub nem a sobreamostragem para compensar o

desbalanço da amostra. Apesar disso, nas próximas sprints, pretende-se coletar mais dados, com o intuito de suavizar essas assimetrias.

4.6. Análise Temporal

Este primeiro gráfico apresenta a evolução anual da emissão de CO_2 na Amazônia Legal entre os anos de 2000 e 2023.

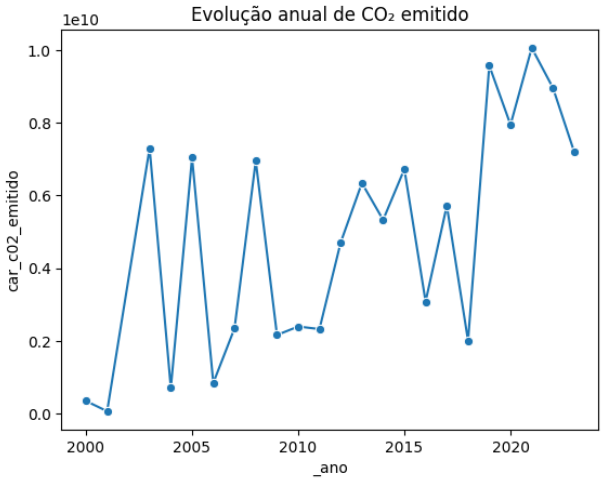

Figura 14 – Evolução anual do CO_2 emitido

Há um padrão bastante oscilante ao longo do período, mas com uma tendência geral de crescimento, principalmente a partir de 2011. Nos primeiros anos da série, entre 2000 e 2003, os valores de emissão eram bastante oscilantes. Em 2003, há um pico, ultrapassando 7 bilhões de toneladas de CO_2 emitidas, seguido por uma queda acentuada em 2004, continuando com altas oscilações até 2010. A partir de 2011, nota-se um crescimento mais contínuo e consistente nas emissões, com menores quedas abruptas e mais anos consecutivos de aumento. Entre 2013 e 2017, os valores se mantêm em patamares intermediários, entre 5 e 7 bilhões de toneladas, com leves oscilações. A partir de 2018, no entanto, há um salto significativo, com os anos de 2019 e 2021 registrando os maiores valores da série, próximos ou acima de 10 bilhões de toneladas de CO_2 emitidos. Após esse pico, observa-se uma leve queda, mas os valores permanecem bastante altos até 2023.

Este outro gráfico mostra como a área queimada na Amazônia Legal evoluiu entre os anos de 2000 e 2023.

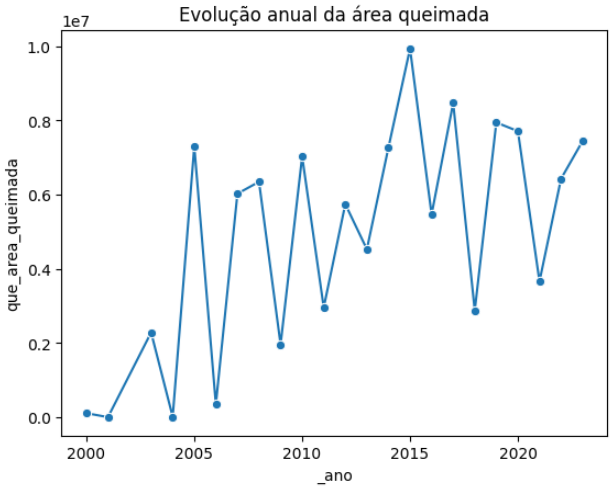

Figura 15 – Evolução anual da área queimada

É possível perceber uma tendência de aumento ao longo do tempo, com várias oscilações de um ano para o outro, que podem ter causas climáticas, políticas ambientais e econômicas. No começo, entre 2000 e 2003, os números eram mais baixos, mas a partir de 2004 já começa um crescimento forte, mesmo com altos e baixos. O ponto mais crítico foi em 2015, quando mais de 10 milhões de hectares foram queimados. Depois disso, os valores continuaram altos, variando bastante a cada ano. Entre 2020 e 2023, houve uma queda seguida de novo aumento. Mesmo nos anos a partir de 2010 que houve queda, os valores ainda são muito maiores do que os do começo da série.

Por fim, tem-se a sobreposição das séries mensais de CO_2, área queimada e quantidade de focos.

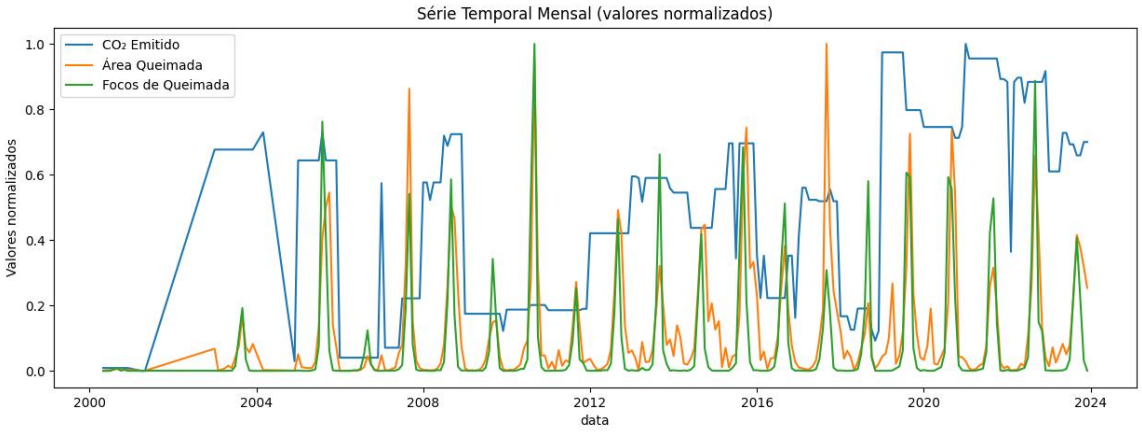

Figura 16 – Série temporal mensal de CO_2 emitido, área queimada e focos

Visualmente, há correlação entre a quantidade de área queimada e a quantidade de focos, haja vista que muitos de seus picos nas séries temporais coincidem. Apesar disso, é nítido que os

picos menores de área queimada não costumam ser acompanhados por picos menores da curva de focos, que se mantém plana. Isso pode estar relacionado ao fato de que grandes focos, embora não numerosos, produzam vasta área de queima.

Quanto às emissões de carbono, a imputação dos valores anuais a todos os meses parece ter introduzido demasiada distorção, de modo que a série de CO_2 mantenha-se em alta grande parte do tempo, mesmo que desacompanhada pelas demais. Diante disso, para evitar interpretações enganosas, deve-se procurar dados propriamente mensais sobre o tópico, ainda que disponíveis em um intervalo mais curto que o escopo do projeto.

4.7. Análise Geoespacial

Neste gráfico, quanto maior a intensidade da cor, maior é a área queimada em cada estado.

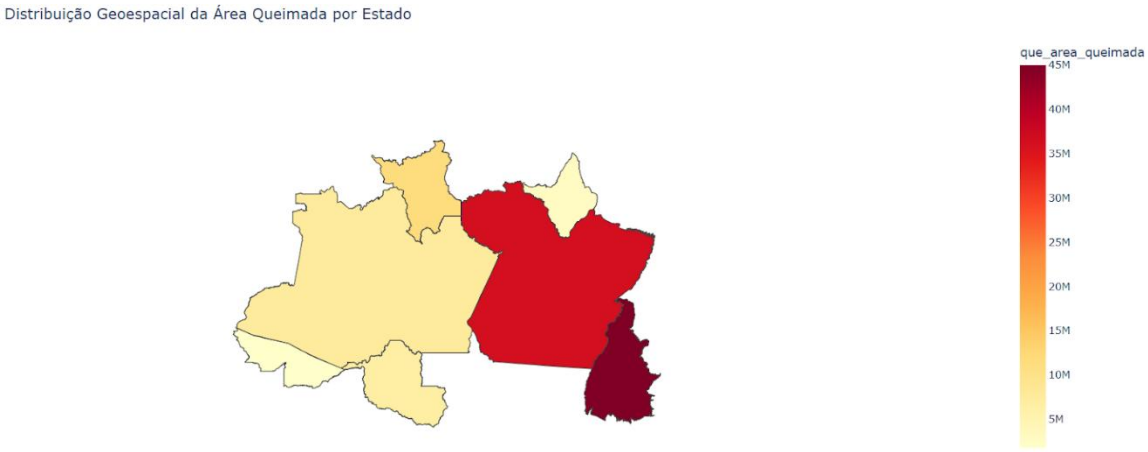

Figura 17 – Área queimada acumulada por estado

Observa-se que o Tocantins, seguido do Pará, apresenta as maiores extensões de áreas queimadas. Em contraste, o Acre e o Amapá registram as menores quantidades. Já os estados do Amazonas, Rondônia e Roraima situam-se em uma posição intermediária. Isso é condizente com a realidade, pois o Pará e o Tocantins fazem parte da chamada Fronteira Agrícola do Brasil, onde há intensa expansão de áreas para agricultura e pastagens. O fogo é usado como uma técnica barata para desmatar e "limpar" o terreno. Além disso, o Tocantins enfrenta anualmente um período de estiagem que se inicia em maio, caracterizado por altas temperaturas e baixa umidade relativa do ar. Essas condições tornam a vegetação mais suscetível ao fogo. Ademais, no Pará há grandes áreas de florestas sendo invadidas ilegalmente para fins de especulação fundiária, com o uso do fogo para marcar territórios e abrir áreas. Por fim, vale ressaltar que ausência de dados sobre Maranhão e Mato Grosso não implica que esses estados não tenham áreas queimadas.

4.8. Insights Importantes

A partir das análises acima, constata-se que, ora por subrepresentação (MA e MT), ora por imputação (emissões mensais de CO_2), distorções e enganos podem ser produzidos, como o Maranhão não ter área queimada. Apesar disso, o problema mais premente parece ser o da emissão de carbono, a qual, além de variável-alvo, mostrou suas distorções diante da discrepância entre a correlação com que_area_queimada antes e depois da agregação por estado. Ademais, algumas correlações mostraram-se surpreendentemente baixas, a saber:

- **`que_area_queimada`** ↔ **`que_focos_qtd`** = 0.48;
- **`que_area_queimada`** ↔ **`cli_temp_ar_med`** = 0.03.

Tal observação revela o fato de que o senso comum pode ser enganoso em algumas situações: com efeito, um pequeno número de grandes focos pode ser mais impactante que um grande número de pequenos. Por outro lado, expectativas confirmaram-se diante das fortes correlações entre variáveis "cognatas", como **`cli_umid_rel_min_med`** e **`cli_umid_rel_min_min`**, o que pode demandar seleção de *features* no ciclo interno de pesquisa com o intuito de evitar a *Curse of Dimensionality*.

Por fim, a análise geoespacial, que permitiu relacionar a quantidade de área queimada à expansão da fronteira agrícola, aponta que não apenas os dados climático-ambientais podem ser úteis, mas também os socioeconômicos. Diante do exposto, conclui-se, em suma, que o universo sugerido pelos dados é um mero reflexo da realidade, que tenderá a ser mais ou menos fiel conforme a qualidade e a limpeza do espelho.

5. Resultados Parciais

A seção a seguir apresenta os resultados parciais obtidos pelos modelos de regressão aplicados à tarefa de previsão das emissões mensais de dióxido de carbono (CO_2) na Amazônia Legal. Para contextualizar adequadamente os desempenhos obtidos, é fundamental conhecer as características estatísticas da variável-alvo `car_co2_emitido`, cuja distribuição influencia diretamente a interpretação dos erros preditivos. A tabela abaixo exibe dois indicadores descritivos centrais dessa variável: o desvio padrão, que representa a dispersão média dos valores em torno da média, e a variância, que expressa a magnitude absoluta da variabilidade dos dados em unidades ao quadrado.

Indicador	Valor
Desvio padrão (toneladas)	$1{,}30 \times 10^{8}$
Variância (toneladas²)	$1{,}69 \times 10^{16}$

Tabela 6 – Medidas estatísticas do alvo

5.1. Regressão Linear Múltipla

A Regressão Linear Múltipla foi adotada após a constatação de que a variável `que_area_queimada`, isoladamente, não apresentava capacidade preditiva satisfatória em relação às emissões de CO_2 na atmosfera. A aplicação de uma Regressão Linear Simples com essa única variável resultou em um coeficiente de determinação (R^2) negativo, evidenciando desempenho inferior ao de um modelo trivial que prediz a média da variável-alvo para todos os casos. Adicionalmente, as métricas de erro, como o Erro Absoluto Médio (MAE) e o Erro Quadrático Médio (MSE), apresentaram valores significativamente superiores à variância da variável dependente, o que reforça o baixo ajuste do modelo e indica que a variabilidade das emissões não foi adequadamente capturada.

Diante disso, optou-se por uma modelagem múltipla, na qual se buscou estimar a emissão mensal de CO_2 como função conjunta de diversos fatores ambientais e indicadores de queimadas. Foram utilizadas todas as variáveis remanescentes no *dataset*, com exceção daquelas removidas por apresentarem alta multicolinearidade, condição que compromete a estabilidade e a interpretabilidade dos coeficientes estimados. As demais variáveis — incluindo indicadores climáticos como temperatura média, umidade mínima média e velocidade média do vento, além de medidas diretas relacionadas às queimadas — foram mantidas e utilizadas na modelagem.

Conforme mencionado na primeira seção, o particionamento da base de dados foi realizado na proporção 70% para treinamento e 30% para teste, permitindo a avaliação do modelo em dados não utilizados no ajuste. O modelo final apresentou desempenho substancialmente superior ao da regressão simples, evidenciando maior aderência aos dados e melhor capacidade explicativa.

5.1.1. Desempenho do Algoritmo

Com o modelo de Regressão Linear Múltipla, foram obtidas as seguintes métricas de desempenho:

Métrica	Valor
R^2	0,9138
MAE (toneladas)	$2{,}77 \times 10^7$
MSE (toneladas²)	$1{,}74 \times 10^5$

Tabela 7 – Métricas da Regressão Linear Múltipla

O desempenho preditivo do modelo pode ser considerado elevado, uma vez que o valor de R^2 indica que aproximadamente 91% da variância nas emissões de CO_2 foi explicada pelo conjunto de variáveis utilizadas, como área queimada e fatores climáticos. No entanto, o erro absoluto médio (MAE) ainda é significativo — da ordem de 27,7 milhões de toneladas — o que sugere a influência de valores extremos ou a necessidade de maior refinamento na seleção e transformação das variáveis. O valor relativamente baixo do MSE, por sua vez, reforça a necessidade de avaliar a escala e a consistência dos dados empregados.

5.1.2. Análise das Variáveis

A análise da importância das variáveis no modelo de Regressão Linear Múltipla foi realizada com base nos coeficientes estimados e nos respectivos p-valores obtidos pelo pacote `Statsmodels`. Como os dados foram previamente padronizados com o `StandardScaler`, os coeficientes da regressão passaram a representar o impacto relativo de cada variável na predição da emissão de CO_2, permitindo uma comparação direta entre variáveis com diferentes escalas originais. Nesse contexto, as variáveis com maior magnitude absoluta nos coeficientes indicam maior influência no modelo, desde que também apresentem significância estatística.

A significância das variáveis pode ser visualizada no gráfico abaixo, que indica o grau de confiança de que cada variável realmente influencia o modelo, sendo que quanto maior a barra, mais relevante é a variável segundo os testes estatísticos.

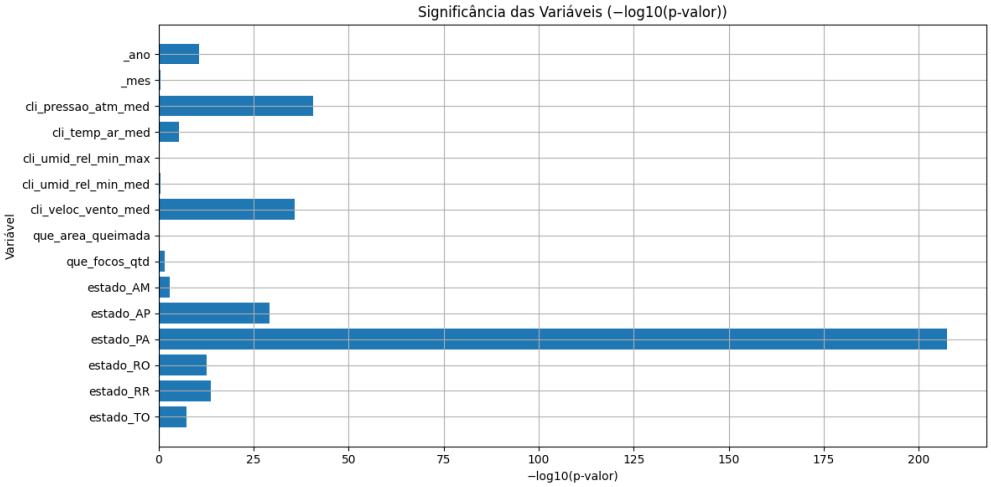

Figura 18 – Significância das variáveis da Regressão Linear Múltipla

Os resultados mostram que algumas variáveis se destacam de maneira expressiva. A variável indicadora **`estado_PA`** apresentou o maior coeficiente positivo, sugerindo que, estando no estado do Pará, há uma forte associação com maiores níveis de emissão de CO_2, em comparação com o estado de referência (automaticamente excluído na codificação *dummy* para evitar multicolinearidade). Já **`estado_AP`** apresentou um coeficiente altamente negativo, indicando uma relação inversa, ou seja, estar no Amapá está associado a níveis mais baixos de emissão. Ambas as variáveis possuem p-valores próximos de zero, o que confirma sua relevância estatística.

Entre as variáveis contínuas, **`cli_veloc_vento_med`** e **`cli_temp_ar_med`** também se destacam. A primeira tem um coeficiente positivo elevado, o que sugere que o aumento da velocidade média do vento está associado ao aumento da emissão de CO_2, enquanto a segunda apresenta um coeficiente negativo considerável, indicando que temperaturas médias do ar mais altas tendem a reduzir a emissão, segundo a relação capturada pelo modelo. Assim como as variáveis anteriores, ambas apresentam p-valores extremamente baixos, reforçando que seus efeitos são estatisticamente significativos.

Além dessas, outras variáveis como **`cli_pressao_atm_med`**, **`_ano`** e **`estado_TO`** também demonstraram significância estatística, embora com coeficientes de menor magnitude, o que indica que, embora contribuam para o modelo, seu impacto relativo é mais discreto. Por outro lado, variáveis como **`_mes`**, **`cli_umid_rel_min_max`**, **`cli_umid_rel_min_med`** e **`que_area_queimada`** apresentaram p-valores altos, sugerindo que sua contribuição para a explicação da variável alvo não é estatisticamente relevante no contexto do modelo ajustado.

Portanto, com base nos coeficientes padronizados e na significância estatística, conclui-se que as variáveis **`estado_PA`**, **`estado_AP`**, **`cli_veloc_vento_med`** e **`cli_temp_ar_med`** são as que

mais impactam o modelo de Regressão Linear Múltipla, exercendo maior influência na predição da emissão de CO_2.

5.1.3. Interpretação dos Resultados

O modelo de Regressão Linear Múltipla aplicado neste estudo apresentou um desempenho satisfatório frente ao objetivo de prever as emissões de CO_2 na Amazônia Legal com base em variáveis ambientais e indicadores de queimadas. O coeficiente de determinação (R^2) de 0,9138 indica que o modelo foi capaz de explicar mais de 91% da variabilidade observada nas emissões, o que representa um grau elevado de aderência aos dados.

Embora o desempenho geral tenha sido bom, o Erro Absoluto Médio (MAE) de $2{,}77 \times 10^7$ toneladas revela que o modelo ainda apresenta desvios consideráveis em valores absolutos, especialmente para observações com valores muito altos de emissão. Isso sugere a presença de *outliers* ou de comportamentos extremos que não são totalmente capturados pela estrutura linear do modelo. O Erro Quadrático Médio (MSE) de $1{,}74 \times 10^5$ toneladas², por outro lado, reforça que a maior parte das previsões encontra-se relativamente próxima dos valores reais.

Esses resultados atendem parcialmente às expectativas do estudo: confirmam que variáveis como área queimada, focos de incêndio e condições climáticas possuem alto poder explicativo sobre as emissões de CO_2, alinhando-se com a hipótese central de que os incêndios florestais são os principais responsáveis pelas variações de emissão atmosférica na região. No entanto, o desempenho limitado em termos de erro absoluto aponta para a necessidade de considerar, em trabalhos futuros, modelos não-lineares ou transformações na variável alvo, a fim de aumentar a precisão preditiva especialmente em cenários extremos.

Em termos práticos, os resultados implicam que ações de controle de queimadas poderiam ter impacto direto e mensurável na redução de emissões, o que reforça a importância de políticas públicas de preservação e monitoramento ambiental na região da Amazônia Legal.

5.1.4. Visualização dos Resultados

Para validar os resultados obtidos com o modelo de regressão linear múltipla, foram geradas diversas visualizações que ajudam a entender a qualidade das previsões, o comportamento dos resíduos (erros) e a adequação aos pressupostos do modelo.

Previsões vs. Valores Reais

A comparação direta entre os valores reais de emissão de CO_2 e os valores previstos pelo modelo foi representada por gráficos de dispersão. Quanto mais os pontos se aproximam da diagonal imaginária (linha ideal), melhor é a capacidade do modelo de reproduzir os dados observados.

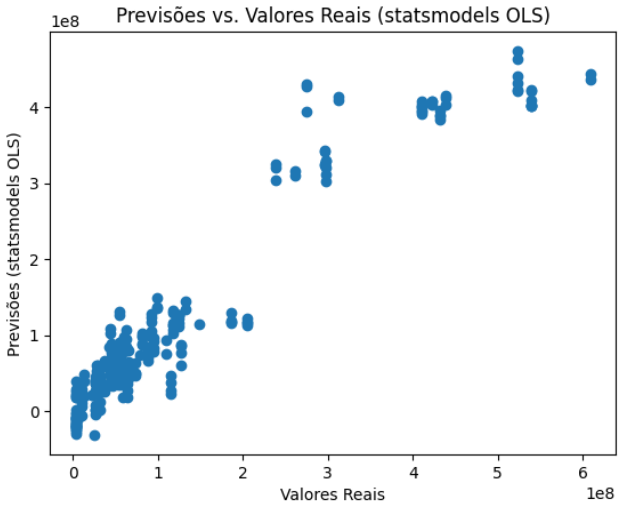

Figura 19 – Previsões *vs* valores reais para a Regressão Linear Múltipla

Distribuição dos Resíduos

Em seguida, foi analisada a distribuição dos resíduos (diferença entre o valor real e o previsto). O ideal é que esses resíduos estejam centrados em zero e distribuídos simetricamente — o que indicaria ausência de viés sistemático nas previsões. O histograma mostrou esse comportamento geral, com pequenas assimetrias nas caudas, provavelmente causadas por *outliers*.

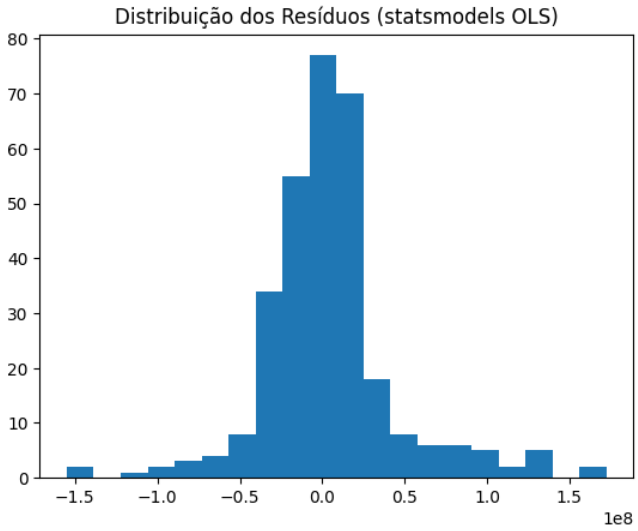

Figura 20 – Histograma dos resíduos para a Regressão Linear Múltipla

Resíduos vs. Valores Ajustados

Por fim, foram analisados os gráficos de resíduos em função dos valores ajustados (preditos). Essa visualização é essencial para verificar a homocedasticidade, ou seja, se a variância dos erros se mantém constante em toda a faixa de valores previstos. Como pode ser visto no gráfico,

os resíduos se distribuíram de forma relativamente aleatória em torno da linha horizontal ($y = 0$), o que sugere que esse pressuposto foi razoavelmente atendido. Contudo, uma leve dispersão nos extremos indica que o modelo pode ser sensível a valores muito altos de emissão — o que é coerente com os outliers identificados anteriormente.

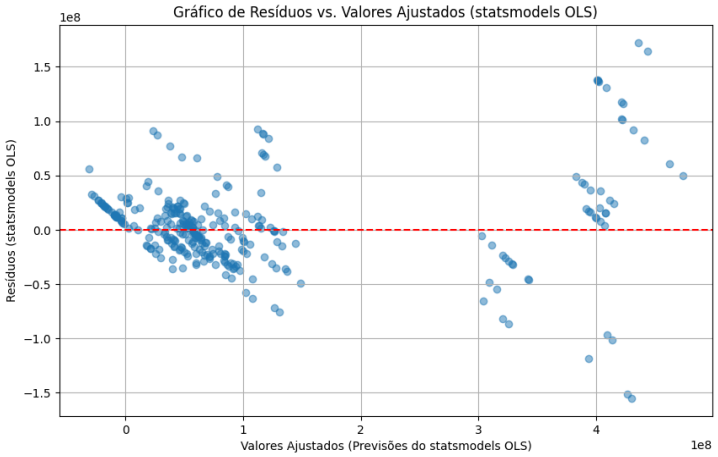

Figura 21 – Resíduos *vs* valores ajustados para a Regressão Linear Simples

Conclusão da Visualização

As visualizações apresentadas reforçam que o modelo de regressão linear múltipla se comporta de maneira estável, com bom poder preditivo e sem grandes desvios estruturais. As análises gráficas, combinadas com as métricas de desempenho, sustentam a validade dos resultados obtidos e apontam para um modelo estatisticamente adequado à proposta do estudo.

5.2. *Árvore de Regressão*

Escolhemos testar o modelo de Árvore de Decisão para Regressão, implementado via `DecisionTreeRegressor` da biblioteca *sklearn*. A escolha desse modelo se deu por vários motivos: além de ser naturalmente interpretável, permitindo a visualização das regras de decisão e a identificação das variáveis mais relevantes para a predição, as Árvores de Decisão apresentam flexibilidade para capturar relações não lineares e interações entre variáveis, o que pode ser uma limitação em modelos lineares tradicionais.

Antes da implementação da Árvore de Decisão, foi realizada uma análise utilizando a Regressão Linear Múltipla. Embora esse modelo tenha apresentado resultados satisfatórios, observou-se que ele possui limitações na captura de relações não lineares e de interações mais complexas entre as variáveis preditoras. Por esse motivo, optou-se por testar a Árvore de Decisão, que, além de ser mais flexível para modelar diferentes distribuições e padrões nos dados, pode identificar interações entre variáveis e fornecer uma maior interpretabilidade ao processo de tomada de decisão. Assim, buscou-se aprimorar o desempenho do modelo e obter uma compreensão mais aprofundada dos fatores que influenciam a variável alvo.

Para a codificação, como o modelo de Árvore do *sklearn* exige variáveis numéricas, as variáveis categóricas (como o estado) foram convertidas via codificação *one-hot* (`pd.get_dummies`), que transforma cada categoria em uma coluna binária, permitindo que o modelo utilize essas informações.

Além disso, para o particionamento, o dataset foi dividido em conjuntos de treino e teste na proporção de 70% para treino e 30% para teste, garantindo que a avaliação do modelo seja feita em dados não vistos durante o treinamento.

A fim de garantir confiabilidade dos resultados, foi feita validação cruzada *K-Fold* (K=5). Isso significa que o conjunto de treino é dividido em 5 partes (*folds*), de modo que, em cada iteração, uma parte é usada como validação e as demais como treino. Dessa forma, permite uma avaliação mais robusta do desempenho do modelo, reduzindo o risco de *overfitting* e fornecendo uma estimativa mais realista da performance em dados não vistos. A função `cross_val` foi implementada para automatizar esse processo, retornando os modelos treinados e as métricas de cada *fold*.

Após a validação cruzada, foi adotado o modelo que tenha maximizado o Coeficiente de Determinação, por ser uma métrica facilmente interpretável. Quanto às demais métricas, por serem baseadas na noção direta de distância (como MAE e MSE), seria mais difícil determinar a partir de qual faixa de erro o modelo poderia ser considerado satisfatório, especialmente sem um referencial claro para o domínio dos dados. Após a versão com maior coeficiente de determinação ser selecionada, foram realizadas previsões sobre os dados de teste. As métricas calculadas nesse conjunto refletem a capacidade do modelo de generalizar para dados realmente não vistos, simulando o desempenho em um cenário real de aplicação. Assim, a avaliação no conjunto de teste complementa a análise da validação cruzada, fornecendo uma visão mais completa e confiável sobre a performance do algoritmo.

5.2.1. Desempenho dos Algoritmo

Para a avaliação do desempenho do algoritmo, foram utilizadas três métricas principais: Coeficiente de Determinação, Erro Quadrado Médio e Erro Absoluto Médio. O resultado das métricas utilizando a validação cruzada foi o seguinte:

Fold	R^2	MAE (toneladas)	MSE (toneladas²)
0	0.9759	$5{,}12 \times 10^{6}$	$5{,}28 \times 10^{14}$
1	0.9991	$9{,}70 \times 10^{5}$	$1{,}04 \times 10^{13}$
2	0.9980	$1{,}00 \times 10^{6}$	$3{,}18 \times 10^{13}$
3	0.9953	$2{,}41 \times 10^{6}$	$1{,}03 \times 10^{14}$
4	0.9810	$3{,}01 \times 10^{6}$	$2{,}50 \times 10^{14}$
Média	0.9899	$2{,}50 \times 10^{6}$	$1{,}85 \times 10^{14}$

Tabela 8 – Métricas da validação cruzada para a Árvore de Decisão

O resultado das predições de teste (externo), utilizando o modelo com maior coeficiente de determinação:

R^2	MAE (toneladas)	MSE (toneladas²)
0,9984	$1{,}29 \times 10^{6}$	$3{,}16 \times 10^{13}$

Tabela 9 – Métricas de teste externo para a Árvore de Decisão

As métricas do modelo foram muito boas, com coeficiente de determinação (R^2) quase igual a 1, como se o modelo fosse perfeito. No entanto, esse desempenho levantou suspeitas de que possa estar ocorrendo *overfitting*, ou seja, o modelo se ajusta excessivamente bem aos dados de treinamento, mas perde capacidade de generalização quando exposto a novos dados. Uma possível causa para isso é a baixa heterogeneidade do *dataset*, já que, como só havia dados de emissões anuais de CO_2 (variável alvo), esses valores foram imputados igualmente nos registros mensais, repetindo-se conforme o ano correspondente. Isso fez com que os dados perdessem variabilidade temporal, tornando mais fácil para o modelo aprender padrões artificiais que não necessariamente refletem a realidade das emissões mensais. Como consequência, o modelo pode parecer muito preciso nas métricas avaliadas, mas na prática não ser capaz de generalizar bem para dados reais com maior variação dentro de um mesmo ano.

5.2.2. Análise das Variáveis

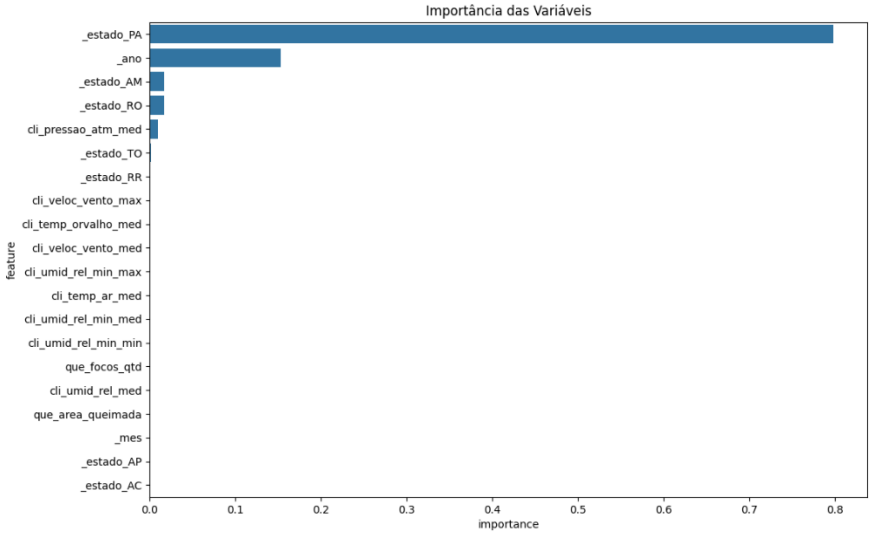

Figura 22 – Importância das variáveis para a Árvore de Decisão

O objetivo do projeto é estimar a quantidade de gás carbônico (CO_2) emitida em um estado, em determinado mês, com base em fatores climáticos, na quantidade de área queimada e no número de focos de incêndio. Inicialmente, esperava-se que as variáveis relacionadas às queimadas e aos fatores climáticos fossem as mais relevantes para a predição. No entanto, ao analisar o gráfico de importância das variáveis gerado pelo modelo de árvore de regressão, observou-se um resultado

inesperado: variáveis como temperatura, umidade, vento, quantidade de área queimada e número de focos de queimada apresentaram importância próxima de zero.

Por outro lado, variáveis categóricas como o estado do Pará (com a maior importância, superior a 80%), o ano, o estado do Amazonas, o estado de Rondônia e, em menor grau, a pressão atmosférica média, foram as que mais influenciaram o modelo.

Esse comportamento reforça a hipótese de que o modelo está capturando padrões específicos associados aos estados e aos anos, em vez de aprender relações generalizáveis entre os fatores ambientais e de queimada com as emissões de CO_2. Isso pode ser um indício de *overfitting*, possivelmente causado pela baixa variabilidade dos dados e pela forma como os valores mensais de emissão de gás carbônico foram imputados com base nas emissões anuais.

5.2.3. Interpretação dos Resultados

Os resultados obtidos com o modelo de árvore de regressão revelaram um alto desempenho com base nas métricas avaliadas, principalmente o Coeficiente de Determinação, que apresentou valores próximos de 1 tanto na validação cruzada quanto no conjunto de teste. À primeira vista, isso indicaria que o modelo é capaz de prever com alta precisão a quantidade de CO_2 emitida, atendendo ao objetivo do estudo.

Entretanto, uma análise mais crítica dos resultados aponta para possíveis problemas quando se considera o objetivo principal do estudo: estimar, em determinado mês e estado, a emissão de CO_2 com base em fatores climáticos e de queimada. A expectativa inicial era de que variáveis diretamente relacionadas às queimadas e aos fatores climáticos, como temperatura, umidade, vento, quantidade de área queimada e número de focos de incêndio, tivessem alta influência sobre as emissões de CO_2. Contudo, contrariando essa ideia, essas variáveis apresentaram importância quase nula no modelo. Por outro lado, variáveis categóricas — como os estados (principalmente o Pará, com mais de 80% de importância) —, o ano e, em menor medida, a pressão atmosférica média, foram as que mais influenciaram as previsões do modelo. Dessa forma, há o comprometimento da capacidade do modelo de atingir o objetivo proposto, pois ele não está, de fato, utilizando os fatores ambientais e de queimada como base para suas decisões preditivas.

Essa discrepância sugere que o modelo pode estar se apoiando fortemente em padrões específicos dos dados, especialmente nas categorias dos estados e dos anos, em vez de aprender relações causais ou generalizáveis entre variáveis climáticas e de queimadas e as emissões de CO_2. Isso levanta a hipótese de *overfitting*, ou seja, o modelo está memorizando os dados ao invés de aprender relações para generalização.

Uma provável explicação para esse comportamento está na maneira como os dados foram preparados. Como os dados de emissão de CO_2 (variável alvo) estavam disponíveis apenas em resolução anual, foi necessário imputar os mesmos valores para todos os meses de um mesmo ano em um estado. Esse procedimento, embora necessário devido a indisponibilidade de dados mensais, reduziu significativamente a variabilidade temporal.

Essa baixa variabilidade pode ter induzido o modelo a aprender que determinados estados ou anos estão diretamente associados a certos níveis de emissão de CO_2, sem de fato considerar as condições ambientais e de queimada que motivam essas emissões.

Assim, a fim de que o modelo atinja seu objetivo (estimar, mês a mês e estado por estado, a quantidade de CO_2 emitida com base em informações ambientais e de queimadas) de maneira confiável e robusta, é necessário aprimorar a base de dados, garantindo maior granularidade temporal e diversidade nos registros. Isso pode envolver a obtenção de dados ou estimativas mensais mais precisas de emissão de CO_2 e a integração com fontes alternativas de dados, como a expansão do agronegócio, que utilizam de queimadas da vegetação como meio para a abertura de novas áreas, aumentando a concentração de gás carbônico na região.

5.2.4. Visualização dos Resultados

A seguir, seguem análises de gráficos referentes ao modelo de árvore de regressão, apresentando visualizações claras a fim de facilitar a compreensão dos resultados.

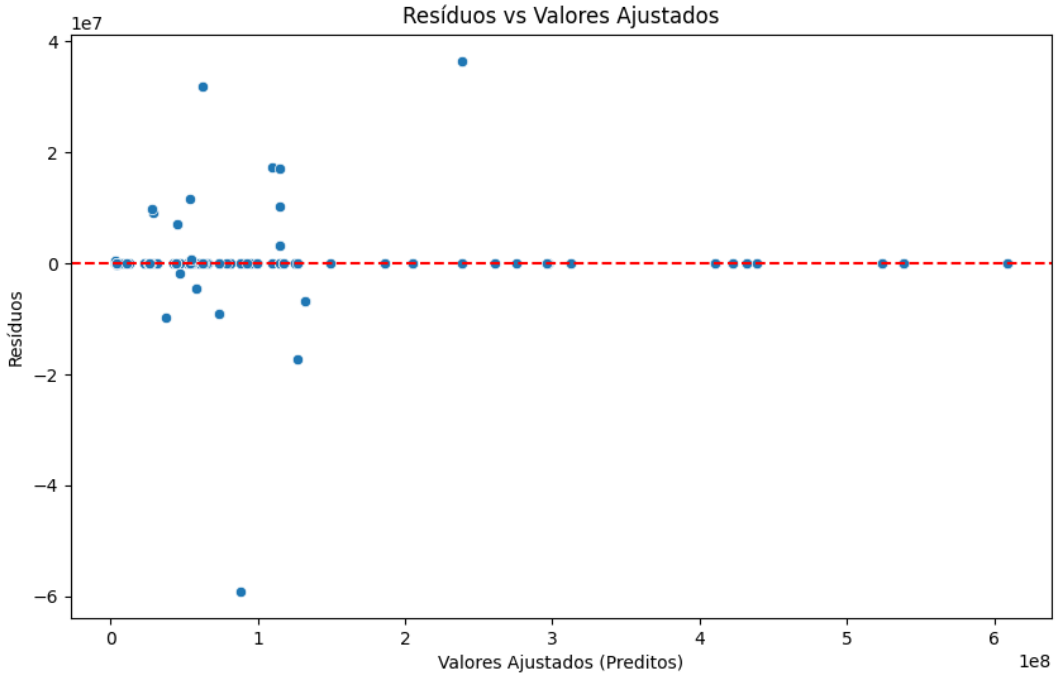

Figura 23 – Resíduos *vs* valores ajustados para a Árvore de Regressão

Resíduos são definidos como a diferença entre os valores reais observados e os valores preditos pelo modelo, indicando o erro de predição. O gráfico de resíduos *versus* valores ajustados (ou preditos) é importante para avaliar a qualidade do ajuste do modelo de regressão. Em um modelo bem ajustado, espera-se que os resíduos estejam distribuídos de forma aleatória em torno da linha zero, sem apresentar padrões de tendências ou comportamentos repetitivos nos dados. No gráfico acima, os valores são em relação a variável alvo, quantidade de gás carbônico emitido. Observa-se que os resíduos estão bastante dispersos quando os valores preditos são baixos, com alguns pontos distantes da linha central, indicando grandes erros de predição ou *outliers*. Entretanto, para valores preditos mais altos, os resíduos estão mais próximos de zero, sugerindo melhor desempenho do modelo nessas situações. Essa assimetria indica que o modelo não está generalizando bem para casos de baixa emissão de CO_2, como citado anteriormente, e pode estar se ajustando demais a padrões específicos do conjunto de dados (possível *overfitting* já comentado).

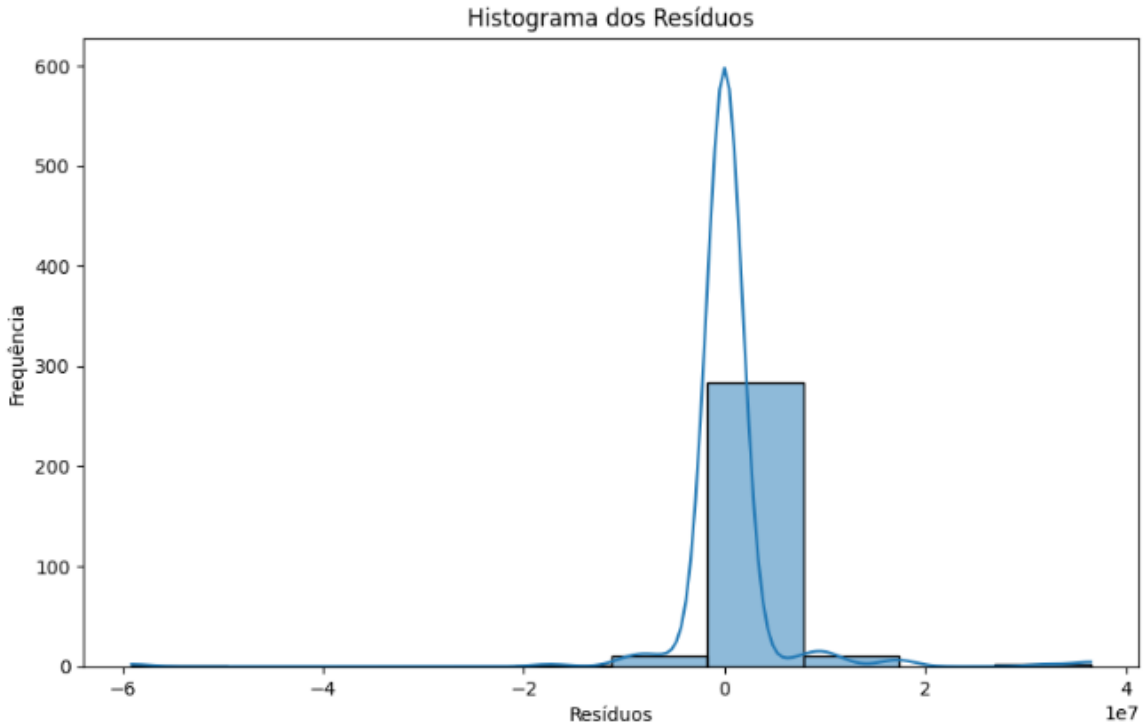

Figura 24 – Histograma de resíduos para a Árvore de Regressão

O gráfico acima mostra o histograma de resíduos do modelo, que mostra a distribuição das diferenças entre os valores reais e os valores preditos. A forma do histograma é um indicativo importante da qualidade do ajuste do modelo. Ele também complementa a análise feita anteriormente no gráfico de resíduos versus valores ajustados. Enquanto o gráfico de dispersão ajuda a verificar a presença de padrões ou curvaturas nos resíduos (indicando erros sistemáticos), o histograma nos ajuda a entender a distribuição e simetria dos erros.

Ao observar o histograma, percebemos que a maioria dos resíduos está concentrada próxima de zero, o que indica que o modelo, na maior parte das vezes, prediz valores bastante próximos dos valores reais. Essa é uma característica desejável em qualquer modelo preditivo, pois mostra que os erros não são grandes nem frequentes.

Além disso, a distribuição dos resíduos tem um formato semelhante ao de uma curva normal (distribuição gaussiana), com um pico bem definido no centro e simetria ao redor do zero. Isso reforça que os erros são, na maior parte dos casos, pequenos e distribuídos de forma aproximadamente simétrica, o que indica que o modelo não está enviesado, ou seja, ele não tende a errar sempre para mais ou para menos.

Dessa forma, a análise do histograma dos resíduos reforça os indícios de que o modelo está bem ajustado aos dados. A concentração dos resíduos em torno de zero e a forma aproximadamente simétrica da distribuição sugerem que os erros cometidos pelo modelo são pequenos, equilibrados e não sistemáticos.

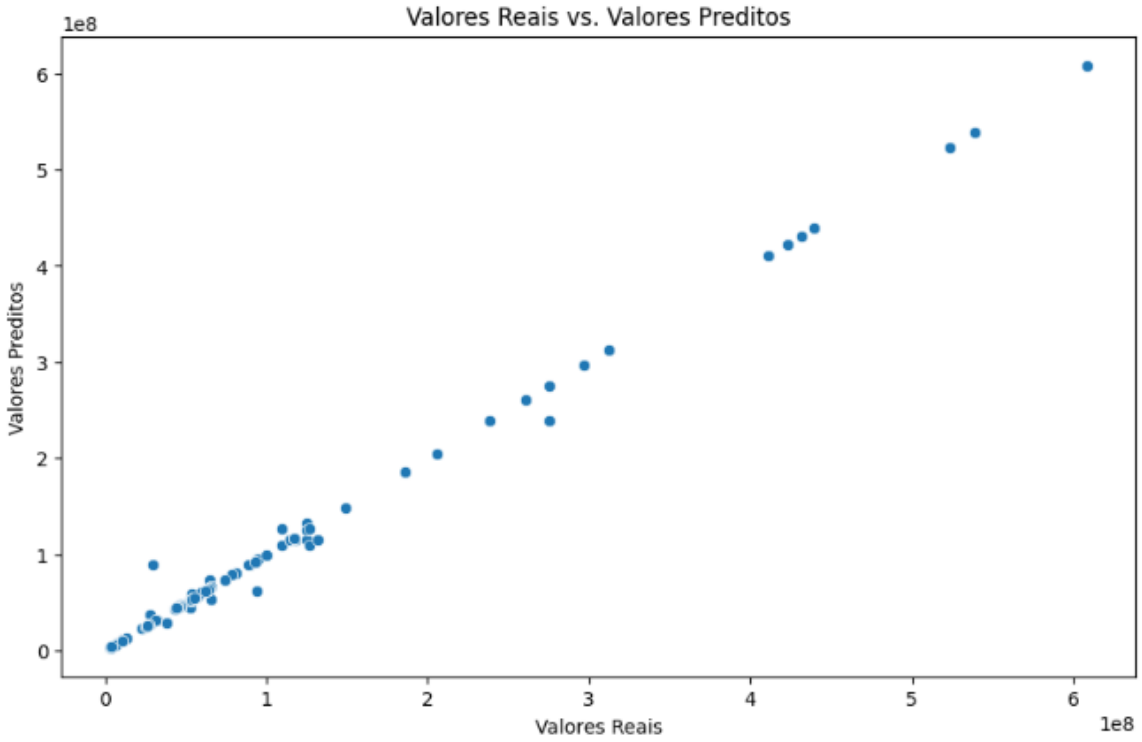

Figura 25 – Valores reais *vs* preditos para a Árvore de Regressão

O gráfico acima representa a dispersão entre os valores reais e os valores preditos pelo modelo de Árvore de Regressão, sendo útil para avaliar visualmente a precisão das previsões. Cada ponto corresponde a uma observação, onde o eixo horizontal indica o valor real das emissões de

CO_2 e o eixo vertical representa o valor previsto pelo modelo. Quanto mais próximos os pontos estiverem da reta $y = x$, melhor será o desempenho do modelo, já que isso indica que os valores preditos se aproximam dos reais.

A análise do gráfico revela que os pontos estão bem concentrados ao redor da reta $y = x$, com pouca dispersão, o que indica alta precisão nas previsões. Esse comportamento corrobora com o coeficiente de determinação obtido ($R^2 = 0{,}9984$), sugerindo que o modelo explica quase toda a variabilidade dos dados, um resultado próximo ao de um modelo perfeito.

Além disso, não são observados *outliers* evidentes nem padrões de curvatura que indiquem erros sistemáticos de predição. Isso sugere que, no conjunto de teste, o modelo se ajustou muito bem aos dados.

Entretanto, apesar do excelente desempenho visual, é importante interpretar esses resultados com cautela. Como discutido anteriormente, o uso de valores replicados para diferentes meses de um mesmo ano e a baixa variabilidade temporal podem ter feito com que ele memorize padrões específicos ao invés de aprender relações generalizáveis.

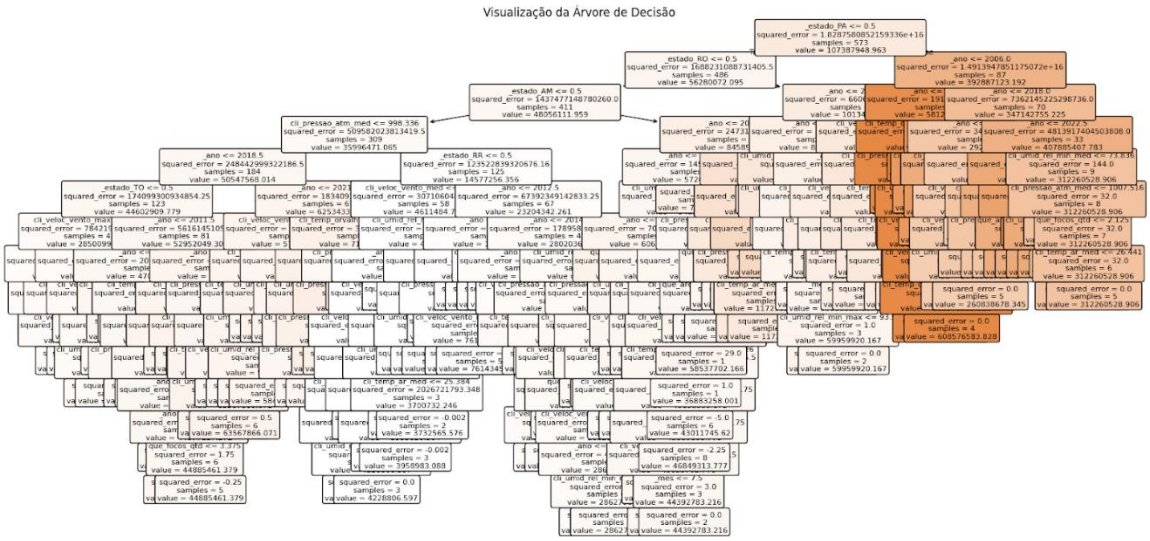

Figura 26 – Diagrama completo da Árvore de Regressão

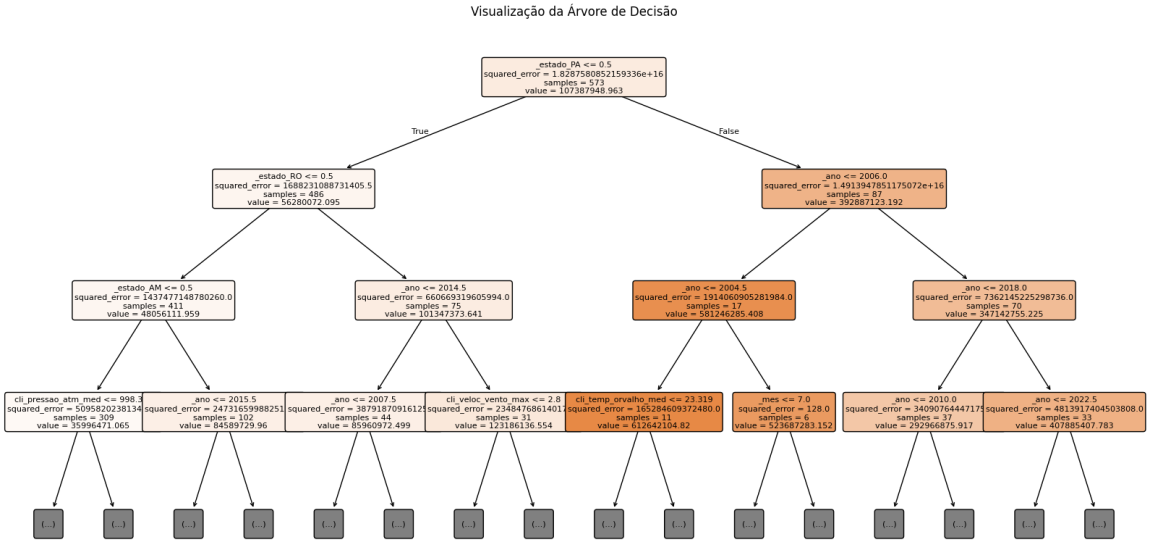

Figura 27 – Diagrama até a profundidade 4 da Árvore de Regressão

O gráfico apresentado é a visualização da Árvore de Decisão gerada pelo modelo de regressão para prever os valores de emissão de CO_2. Cada caixa representa um nó da árvore, onde ocorre uma divisão dos dados com base em uma condição sobre uma variável explicativa (por exemplo, ano, estado, temperatura, umidade, entre outras). O objetivo dessa divisão é segmentar o conjunto de dados em grupos cada vez mais homogêneos em relação à variável resposta (emissão de CO_2). Em cada nó, são mostradas informações como o erro quadrático médio, o número de amostras e o valor médio da variável resposta dentro daquele subconjunto.

A estrutura da Árvore mostra que variáveis como o estado, o ano e características climáticas (pressão atmosférica, temperatura, umidade, velocidade do vento) foram fundamentais para as decisões do modelo. Por exemplo, umas das primeiras divisões são baseadas como o Amazonas, o Pará e Rondônia, o que sugere que a localização geográfica tem grande influência nas emissões. Em seguida, vemos ramificações que se aprofundam com base em variáveis climáticas e temporais, evidenciando que essas características também são altamente relevantes.

Os nós terminais (ou folhas), especialmente aqueles destacados em tons alaranjados, representam os agrupamentos finais de dados para os quais o modelo faz previsões. A intensidade da cor está associada ao valor médio de emissão de CO_2 naquele grupo, quanto mais escuro, maior o valor. Observa-se que, em algumas folhas, o erro quadrático é zero, o que significa que, para aquele subconjunto, o modelo previu exatamente os valores reais, o que pode ocorrer quando há repetição ou pouca variabilidade nos dados daquele grupo.

É interessante destacar também a profundidade da árvore, o que indica que o modelo foi bastante ajustado aos dados (o que chamamos de *overfitting* em alguns casos). Embora o ajuste

aparente seja bom, como vimos nas análises anteriores (gráfico de valores reais vs. preditos e resíduos), árvores muito profundas podem generalizar mal em novos dados, por captarem ruídos ou padrões muito específicos do conjunto de treino.

6. Referências

ANDREAE, M. O.; MERLET, P. Emission of trace gases and aerosols from biomass burning. **Global Biogeochemical Cycles**, v. 15, n. 4, p. 955-966, 2001. Disponível em: https://doi.org/10.1029/2000GB001382. Acesso em: 6 jun. 2025.

ARAGÃO, L. E. O. C.; ANDERSON, L. O.; FONSECA, M. G.; *et al.* 21st century drought-related fires counteract the decline of Amazon deforestation carbon emissions. **Nature Communicatio*ns***, [S.l.], v. 9, p. 536, 2018. Disponível em: https://doi.org/10.1038/s41467-017-02771-y. Acesso em: 06 jun. 2025.

ECOA PUC-RIO. **AMazonizAR**. *Homepage* da instituição, 2023. Disponível em: https://instituto.ecoa.puc-rio.br/amazonizar/. Acesso em: 29 mar. 2025.

G1. **Seca severa no Amazonas já afeta mais de meio milhão de pessoas, aponta Defesa Civil.** G1 Amazonas, 16 out. 2023. Disponível em: https://g1.globo.com/am/amazonas/noticia/2023/10/16/seca-severa-no-amazonas-ja-afeta-mais-de-meio-milhao-de-pessoas-aponta-defesa-civil.ghtml. Acesso em: 06 jun. 2025.

IBGE. **IBGE atualiza mapa da Amazônia Legal**. Agência de Notícias IBGE, 16 jul. 2020. Disponível em: https://agenciadenoticias.ibge.gov.br/agencia-sala-de-imprensa/2013-agencia-de-noticias/releases/28089-ibge-atualiza-mapa-da-amazonia-legal. Acesso em: 4 abr. 2025.

INSTITUTO DE PESQUISA ECONÔMICA APLICADA (IPEA). A floresta amazônica e as mudanças climáticas globais. **Revista Desafios do Desenvolvimento**, Brasília, ano 5, n. 44, 08 jun. 2008. Disponível em: https://www.ipea.gov.br/desafios/index.php?option=com_content&id=2154:catid%3D28. Acesso em: 06 jun. 2025.

INSTITUTO NACIONAL DE METEOROLOGIA (INMET). *Dados históricos*. Brasília: INMET, [s.d.]. Disponível em: https://portal.inmet.gov.br/dadoshistoricos. Acesso em: 2 maio 2025.

INSTITUTO NACIONAL DE PESQUISAS ESPACIAIS (INPE). *Programa Queimadas*. São José dos Campos: INPE, 2025a. Disponível em: https://terrabrasilis.dpi.inpe.br/queimadas. Acesso em: 2 maio 2025.

INSTITUTO NACIONAL DE PESQUISAS ESPACIAIS (INPE). Terrabrasilis: queimadas — situação atual — estatísticas por estados, 2025b. Disponível em: https://terrabrasilis.dpi.inpe.br/queimadas/situacao-atual/estatisticas/estatisticas_estados/. Acesso em: 06 jun. 2025.

INTERGOVERNMENTAL PANEL ON CLIMATE CHANGE (IPCC). Climate change 2021: the physical science basis. Chapter 12 – Climate change information for regional impact and for risk assessment. In: _____. **Sixth assessment report of the Intergovernmental Panel on Climate Change – Working Group I**. Cambridge: Cambridge University Press, 2021. Disponível em: https://www.ipcc.ch/report/ar6/wg1/downloads/report/IPCC_AR6_WGI_Chapter12.pdf. Acesso em: 06 jun. 2025.

KAUFMAN, Y. J.; JUSTICE, C. O.; FLYNN, L.; *et al.* Fire detection from EOS-MODIS. *Journal of Geophysical Research: Atmospheres*, [S.l.], v. 103, n. D24, p. 32215–32238, 1996. Disponível em https://modis-images.gsfc.nasa.gov/_docs/Kaufman%20et%20al.%20(1998c).pdf. Acesso em: 06 jun. 2025.

MAPBIOMAS. *Monitor do Fogo*. São Paulo: Projeto MapBiomas, [s.d.]. Disponível em: https://plataforma.brasil.mapbiomas.org/monitor-do-fogo. Acesso em: 2 maio 2025.

NEPSTAD, D.; SCHWARTZMAN, S.; BAMBERGER, B.; *et al.* Inhibition of Amazon deforestation and fire by parks and indigenous lands. *Conservation Biology*, [S.l.], v. 20, n. 1, p. 65–73, 2006. Disponível em: https://doi.org/10.1111/j.1523-1739.2006.00351.x. Acesso em: 06 jun. 2025.

OBSERVATÓRIO DO CLIMA. **Sistema de Estimativas de Emissões de Gases de Efeito Estufa (SEEG)**. Disponível em: https://plataforma.seeg.eco.br/. Acesso em: 29 mar. 2025.

SAATCHI, S. S.; HOUGHTON, R. A.; DOS SANTOS ALVALÁ, R. C.; SOARES, J. V.; YU, Y. Distribution of aboveground live biomass in the Amazon basin. **Global Change Biology**, Oxford, v. 13, n. 4, p. 816–837, 2007. Disponível em: https://cce.nasa.gov/veg3dbiomass/saatchi-etal_GCB07.pdf. Acesso em: 06 jun. 2025.

SEILER, W.; CRUTZEN, P. J. Estimates of gross and net fluxes of carbon between the biosphere and the atmosphere from biomass burning. **Climatic Change**, [S.l.], v. 2, n. 3, p. 207–247, 1980. Disponível em: https://doi.org/10.1007/BF00137988. Acesso em: 06 jun. 2025.

SISTEMA DE ESTIMATIVAS DE EMISSÕES DE GASES DE EFEITO ESTUFA (SEEG). Relatório anual SEEG: emissões de gases de efeito estufa no Brasil (1970–2022). [S.l.]: Observatório do Clima, 2023. Disponível em: https://oc.eco.br/wp-content/uploads/2023/11/Relatorio-SEEG_gases-estufa_2023FINAL.pdf. Acesso em: 06 jun. 2025.