# Relatório Analítico de Desmatamento

Pedro Mambelli Fernandes

2025-06-30

## Introdução

Este documento detalha a jornada e o processo de desenvolvimento do projeto de análise preditiva de desmatamento para o Desafio Zetta Labs 2025, abordando a metodologia, os resultados e as recomendações estratégicas.

## Metodologia

A análise inicial focou nos dados do PRODES na Amazônia, especificamente no estado do Pará. O objetivo principal era interligar dados socioeconômicos com previsões de desmatamento.

### Validação *Walk-Forward*

Para avaliar a capacidade preditiva dos modelos, foi utilizada a validação *walk-forward*, uma técnica de validação cruzada que divide o conjunto de dados no tempo, treinando o modelo com os dados até o tempo t-1 e testando-o com o dado no tempo t.

### Abordagem Inicial e Desafios

Inicialmente, explorei a possibilidade de prever os Índices de Progresso Social (IPS) com base no desmatamento. Contudo, o IPS de 2024 era a única informação socioeconômica disponível, uma vez que o índice começou a ser compilado, no Brasil, apenas em 2024, e a natureza multifatorial do índice dificultou a previsão precisa utilizando apenas os dados de desmatamento. Além disso, observou-se que modelos lineares não conseguiam explicar adequadamente o fenômeno, como indicado por correlações lineares não muito fortes nas matrizes de correlação, o que é compreensível dada a complexidade do desmatamento.

### Mudança de Abordagem: Análise Temporal e Modelagem Tabular

Diante dos desafios iniciais, a abordagem foi ajustada para uma análise temporal, com foco na modelagem tabular. Isso permitiu incorporar o desmatamento do ano corrente e dos anos anteriores (como t-1 e médias móveis) como features. Dois modelos principais foram selecionados para esta análise: Random Forest e XGBoost, escolhidos pela sua capacidade de lidar com a complexidade do fenômeno do desmatamento, que modelos lineares não conseguiam capturar eficientemente.

## Incorporação de Contexto Político e Novos Dados

A modelagem inicial, mesmo com a clusterização de municípios pelos índices, não apresentou bons resultados para anos de mudanças políticas abruptas, especificamente 2019 e 2023. Uma investigação mais aprofundada revelou que o Programa de Prevenção e Controle do Desmatamento na Amazônia (PPCDAm) foi suspenso em 2019 e retomado em 2023. Os modelos tenderam a subestimar o desmatamento real no período de suspensão (2019-2022) e a superestimá-lo após a retomada (2023), indicando que o efeito do programa (ou fatores correlacionados) não foi imediatamente absorvido pelos modelos mais simples.

Para enriquecer os dados e melhorar a capacidade preditiva dos modelos, foram incorporados os autos de infração do IBAMA de forma anual, incluindo o número de multas e a quantidade de multas relacionadas à categoria "flora" por município e ano. Além disso, os índices principais do IPS foram incluídos diretamente na modelagem, em vez de serem utilizados apenas para clusterização.

## Metodologia SEMMA

O projeto seguiu a metodologia SEMMA (Sample, Explore, Modify, Model, Assess) para garantir uma abordagem estruturada e reprodutível no pipeline de machine learning.

- **Sample (Amostragem)**: Carregamento inicial dos dados de desmatamento anual, infrações do IBAMA e Índice de Progresso Social (IPS) para o estado do Pará, com a geração de resumos estatísticos.
- **Explore (Exploração)**: Análise exploratória dos dados, focando na filtragem geográfica (Pará), identificação e visualização de valores ausentes, e validação da cobertura temporal (a partir de 2008).
- **Modify (Modificação)**: Pré-processamento e engenharia de features. Esta etapa incluiu o carregamento inteligente, filtragem temporal e geográfica, limpeza de dados (padronização de códigos IBGE e tratamento de ausentes), agregações estratégicas (IBAMA por município/ano, interpolação de IPS) e criação de features temporais (lags, médias móveis, e períodos presidenciais).
- **Model (Modelagem)**: Treinamento de modelos XGBoost e Random Forest utilizando uma estratégia de modelagem temporal walk-forward (previsão do próximo ano com base em dados históricos). Foi realizado ajuste bayesiano dos hiperparâmetros com Optuna.
- **Assess (Avaliação)**: Análise comparativa da performance dos modelos, incluindo a evolução do R2 e RMSE ao longo dos anos de teste, importância das features, e a evolução temporal do desmatamento real vs. previsto com contexto político (suspensão do PPCDAm).

## Resultados

Os modelos Random Forest e XGBoost foram avaliados e comparados, com os seguintes resultados médios:

| Modelo | RMSE Médio (km²) | R² Médio |
|---|---|---|
| Random Forest | 22.2 | 0.861 |
| XGBoost | 20.3 | 0.877 |

O XGBoost demonstrou um desempenho ligeiramente superior em termos de R2 médio e RMSE médio.

### Desempenho Temporal dos Modelos

A evolução do R2 e RMSE ao longo dos anos de teste para ambos os modelos é apresentada abaixo:

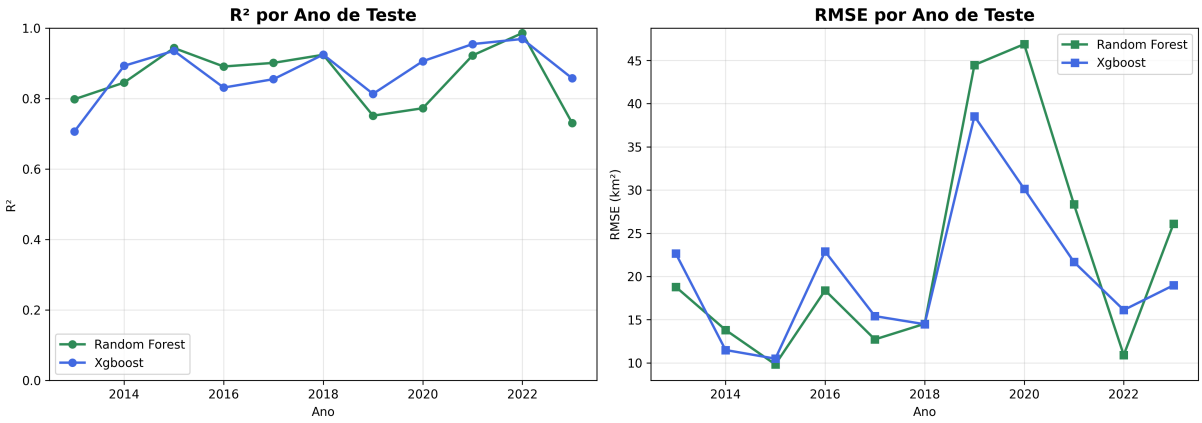


Figura 1: Comparação de Modelos ao Longo do Tempo

**Análise da Performance Temporal:** Os resultados evidenciam padrões distintos na evolução da capacidade preditiva dos modelos ao longo do período de teste (2014-2023). O coeficiente de determinação ($R^2$) apresenta oscilações significativas, com ambos os modelos atingindo performance máxima em 2022 ($R^2$ 0.98), seguida de uma degradação substancial em 2023. Paralelamente, o RMSE demonstra volatilidade considerável, com picos notáveis em 2020-2021, período coincidente com intensificação do desmatamento durante a pandemia e flexibilização das políticas ambientais. A convergência de performance entre Random Forest e XGBoost na maioria dos anos sugere que ambos os algoritmos capturam adequadamente os padrões temporais dos dados, com o XGBoost apresentando ligeira vantagem em estabilidade.

### Evolução Temporal do Desmatamento: Observações Durante Mudanças Políticas

A análise da evolução temporal do desmatamento, comparando os valores reais com as previsões dos modelos, revelou a influência de mudanças políticas, especialmente em relação à suspensão e retomada do PPCDAm.

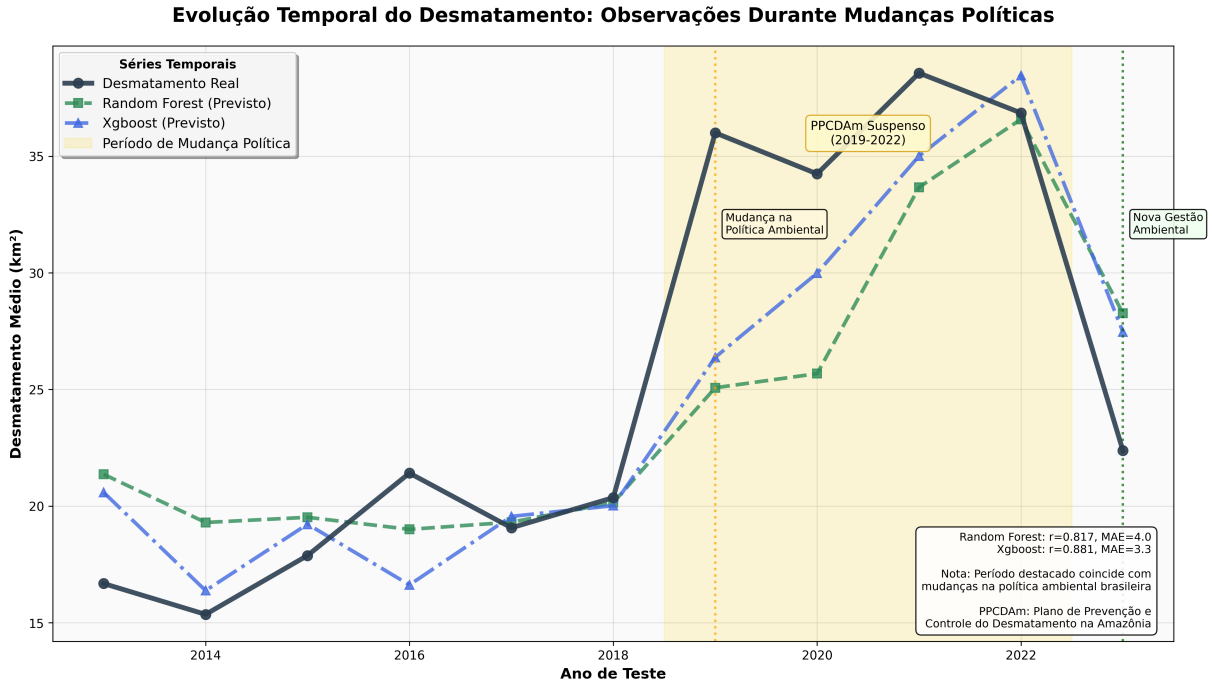


Figura 2: Evolução Temporal: Real vs Previsto

**Análise do Impacto de Políticas Públicas:** O gráfico revela de forma inequívoca o impacto das mudanças na política ambiental brasileira sobre o desmatamento na Amazônia. Durante o período de suspensão do PPCDAm (2019-2022), observa-se uma divergência significativa entre as previsões dos modelos e os valores reais de desmatamento. Os modelos, treinados com dados históricos do programa ativo, sistematicamente subestimaram o desmatamento real, com o erro médio absoluto atingindo 4.0 km² para Random Forest e 3.3 km² para XGBoost. A retomada do programa em 2023 resultou em uma redução abrupta do desmatamento real², demonstrando a eficácia imediata das políticas de controle ambiental. Esta análise confirma a hipótese de que fatores político-institucionais exercem influência determinante sobre os padrões de desmatamento, superando variáveis socioeconômicas tradicionais.

## Performance em Anos Críticos

A performance dos modelos foi analisada especificamente em anos críticos de mudança política (2019 e 2023), onde se observou um desafio na previsão devido à alteração de contextos.

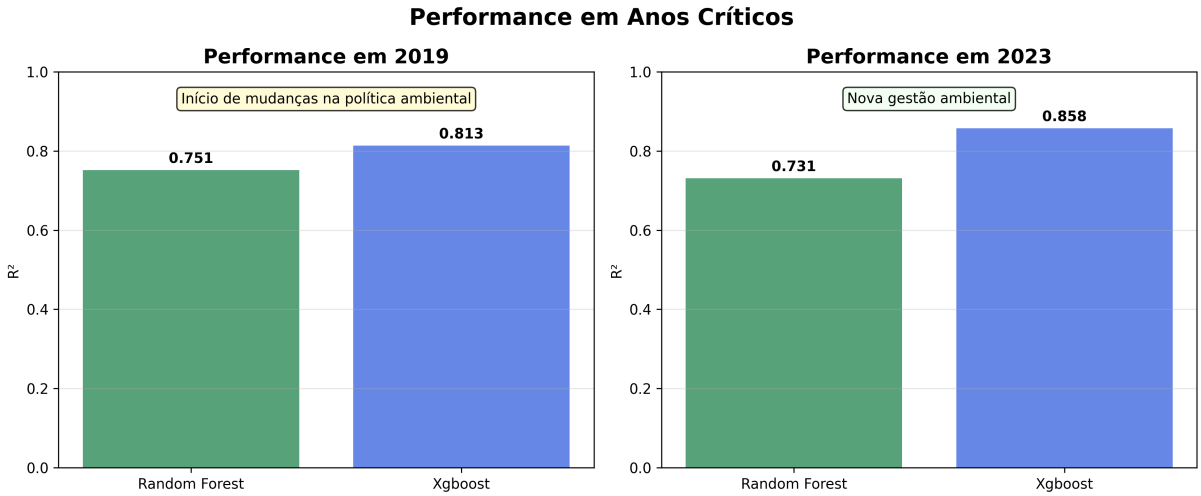


Figura 3: Análise de Anos Críticos

**Análise de Ruptura Política e Performance Preditiva:** A comparação da performance dos modelos nos anos de transição política (2019 e 2023) revela padrões comportamentais distintos dos algoritmos em cenários de mudança estrutural. Em 2019, marco do início das mudanças na política ambiental, o XGBoost demonstrou superioridade com $R^2$ de 0.813 comparado aos 0.751 do Random Forest, sugerindo maior capacidade de adaptação a mudanças iniciais de regime. Em contraste, em 2023, ano de retomada da nova gestão ambiental, ambos os modelos apresentaram performance reduzida (Random Forest: 0.731, XGBoost: 0.858), indicando que mudanças abruptas de política requerem período de adaptação para recalibração dos padrões preditivos. A diferença de performance de 0.127 pontos em favor do XGBoost em 2023 demonstra sua maior robustez a choques estruturais, atributo crítico para aplicações em cenários de instabilidade política.

## Top 10 Features Mais Importantes por Modelo

A importância das features para cada modelo foi avaliada, destacando os fatores que mais influenciam as previsões de desmatamento.

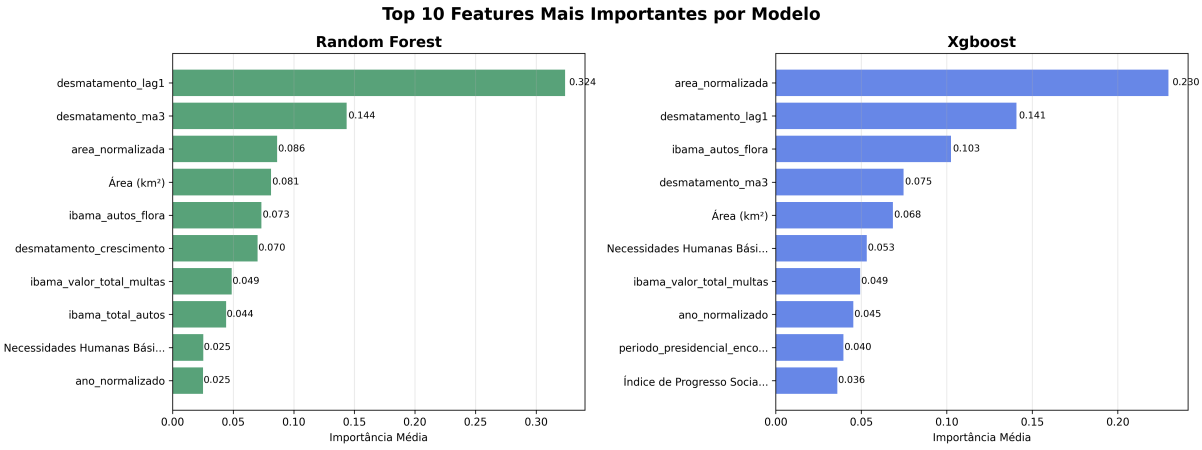


Figura 4: Comparação de Importância das Features

**Análise da Hierarquia de Features Preditivas:** A análise da importância das variáveis

revela convergências e divergências estratégicas entre os algoritmos. Ambos os modelos identificam `desmatamento_lag1` como variável primária (Random Forest: 0.324, XGBoost: 0.141), confirmando a persistência temporal como principal driver preditivo. Entretanto, o XGBoost prioriza `area_normalizada` como feature dominante (0.280), enquanto o Random Forest a posiciona em terceiro lugar (0.086), sugerindo diferentes estratégias de captura de padrões espaciais. Notavelmente, o XGBoost demonstra maior sensibilidade a variáveis de enforcement (`ibama_autos_flora`: 0.103) e indicadores socioeconômicos (`Necessidades Humanas Básicas`: 0.053), enquanto o Random Forest enfatiza features temporais (`desmatamento_ma3`: 0.144).

## Recomendações

Com base nos resultados e na análise do projeto, as seguintes recomendações são propostas:

### 1. Monitoramento Contínuo de Políticas Públicas

É crucial integrar informações sobre políticas públicas e seus períodos de vigência diretamente na modelagem. Modelos mais complexos ou a inclusão de features que capturem o impacto dessas políticas podem melhorar a acurácia em cenários de mudanças abruptas.

### 2. Enriquecimento de Dados Socioeconômicos e Ambientais

A inclusão de dados mais granulares e frequentes sobre o IPS e outras variáveis socioeconômicas, bem como dados mais detalhados sobre as infrações do IBAMA (e.g., tipos específicos de infração, valores aplicados, desfechos dos processos), pode aprimorar significativamente a capacidade preditiva dos modelos.

### 3. Exploração de Modelos Híbridos ou Adaptativos

Para lidar com a não-linearidade e a complexidade do desmatamento, a exploração de modelos híbridos que combinem a robustez de modelos como XGBoost e Random Forest com abordagens de séries temporais pode ser benéfica. Além disso, modelos adaptativos que ajustam seus pesos ou parâmetros em tempo real a novas informações políticas ou ambientais podem ser considerados.

### 4. Análise de Sensibilidade e Cenários

Realizar análises de sensibilidade para entender como as previsões de desmatamento são afetadas por variações nas features mais importantes. A criação de cenários futuros (e.g., com diferentes níveis de fiscalização ou implementação de políticas) pode auxiliar na formulação de políticas públicas mais eficazes.

### 5. Validação Cruzada Temporal Robusta

Continuar utilizando e aprimorando a validação walk-forward para garantir que os modelos sejam robustos e generalizáveis para dados futuros, evitando vazamento de informações.

### 6. Colaboração Multidisciplinar

A complexidade do desmatamento sugere que uma abordagem multidisciplinar, envolvendo especialistas em ciências ambientais, economia e políticas públicas, pode gerar insights mais profundos e features mais relevantes para a modelagem.

## Repositório

O repositório contendo todos os dados, scripts de processamento, notebooks explicativos e um dashboard interativo pode ser encontrado em: https://github.com/uaipedro/zetta-labs-2.git.

No repositório, você encontrará:

- Scripts para todo o pipeline de dados e modelagem (pasta `scripts/`)
- Notebooks detalhados com a metodologia passo a passo (pasta `notebooks/`)
- Dados brutos e processados (pasta `data/`)
- Relatórios, figuras e resultados dos experimentos (pasta `reports/`)
- Um dashboard interativo para exploração dos resultados (`dashboard_app/` e `dashboard.py`)

Para instruções detalhadas de execução, consulte o arquivo `README.md` no próprio repositório.