Relatório de Desenvolvimento: Sistema de Predição de Risco de Incêndio (Fire Risk Predictor)

Otávio Ferracioli Coletti, Samuel Rubens Souza Oliveira e Gustavo Lelli Guirao

[1]Universidade de São Paulo
São Carlos – SP – Brasil

otaviocoletti@usp.com, samuelrubens@usp.br e gustavo.lelli@usp.br
Números USP: 11767796, 11912533, 11918182

1. Introdução, Motivação e Objetivos

A motivação primordial para o desenvolvimento deste projeto é enfrentar a crise crescente de incêndios florestais no Brasil. Em 2024, o país registrou **278.229 focos de calor**, o número mais alto em 14 anos. A Amazônia, bioma mais afetado, teve mais de 140 mil focos, representando um aumento de **77%** em relação a 2023. As consequências são críticas, incluindo perda de biodiversidade, emissão massiva de CO_2, problemas respiratórios e prejuízos econômicos bilionários.

O objetivo principal deste trabalho é **identificar as condições ambientais propícias a incêndios** e utilizar algoritmos de aprendizado de máquina para predição de risco. Foram empregados três algoritmos (Random Forest, MLP e XGBoost) e seus desempenhos foram comparados para determinar a melhor capacidade de generalização.

2. Bases de Dados e Atributos Utilizados

O sistema de predição foi desenvolvido utilizando o **Dataset SISAM** (Sistema de Informações de Saúde Ambiental), que contém aproximadamente **2,6 milhões de registros de focos de calor** desde 2003.

Para treino e validação, foi utilizada uma amostra balanceada de **250.000 registros** (125 mil com ocorrência de incêndio e 125 mil sem). O modelo final foi testado em um conjunto separado de **50.000 registros**.

Os atributos (features) utilizados na modelagem incluíram:

- Coordenadas geográficas: `longitude` e `latitude`.
- Parâmetros atmosféricos e de poluição: `co_ppb`, `no2_ppb`, `o3_ppb`, `pm25_ugm3`, e `so2_ugm3`.
- Condições climáticas: `precipitacao_mmdia`, `temperatura_c`, `umidade_relativa_percentual`, `vento_direcao_grau`, e `vento_velocidade_ms`.
- Variável alvo: `incendio` (binário, 0 ou 1).

Colunas como `data_pas`, `satelite`, `bioma`, `risco_fogo`, entre outras, foram descartadas antes da modelagem.

3. Pipeline de Desenvolvimento e Clusterização

O pipeline de desenvolvimento empregou a clusterização inicial dos dados geoclimáticos para criar modelos especializados por região, seguida pela aplicação de modelos de regressão para a predição.

3.1. Etapa de Clusterização

A clusterização foi realizada para agrupar pontos de monitoramento com base em características ambientais similares, o que permite reconhecer perfis de risco distintos entre as regiões. Como vamos analisar dados de pontos de monitoramento do Brasil inteiro, é necessário identificar a partir de um gráfico de correlação quais condições climáticas mais importam para o desencadeamento ou não de um incêndio. A Figura 1 apresenta a matriz de correlação entre as features, enquanto a Figura 2 mostra a distribuição dos pontos de monitoramento.

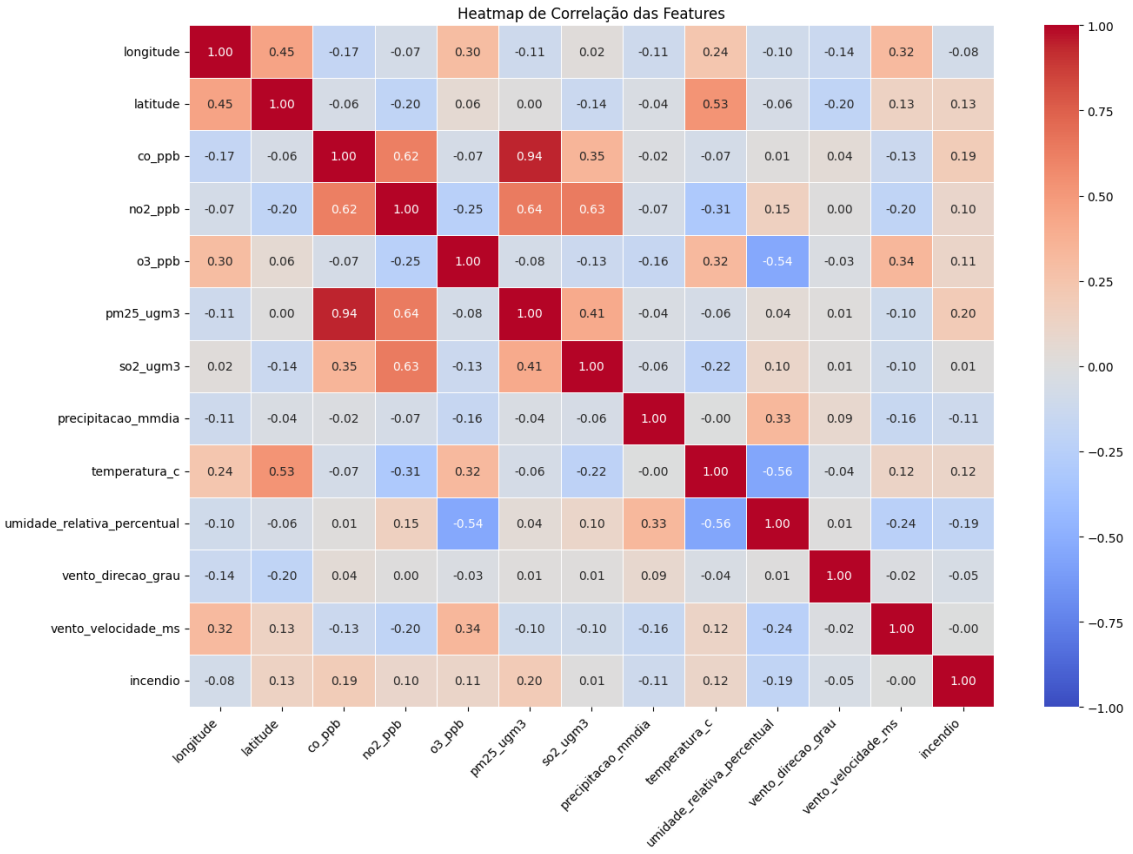

Figure 1. Matriz de correlação entre as features do dataset

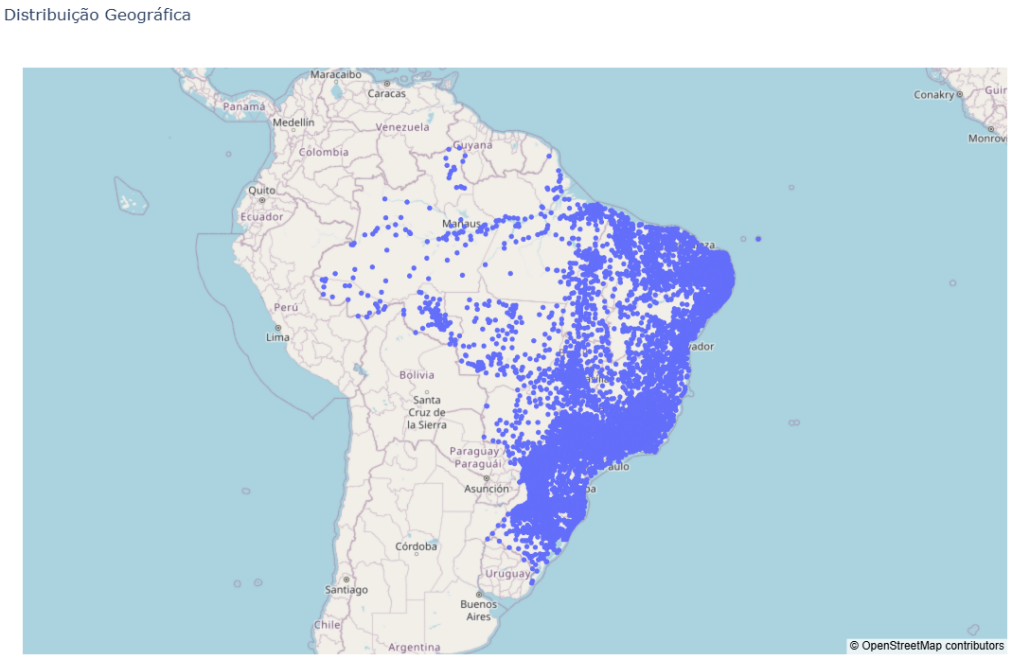

Figure 2. Distribuição dos pontos de monitoramento no Brasil

Com base na análise de correlação, os atributos `temperatura_media`, `co_ppb`, `no2_ppb`, `o3_ppb` e `umidade_relativa_percentual` foram selecionados como os mais importantes para a detecção. A metodologia utilizada na clusterização seguiu os seguintes passos:

1. **Preparação de Dados:** Os dados foram agrupados por `latitude` e `longitude`, calculando-se a média dos atributos relevantes em um período de um ano.

2. **Algoritmo e Parâmetros:** Foi utilizado o algoritmo **K-Means** com o número de grupos fixado em **k=6**.

3. **Features para Clusterização:** As features usadas foram a `temperatura_media`, `co_ppb`, `no2_ppb`, `o3_ppb` e `umidade_relativa_percentual`.

O resultado da clusterização pode ser visto na Figura 4, na qual os clusters gerados são muito semelhantes aos biomas do Brasil, apresentados na Figura 3.

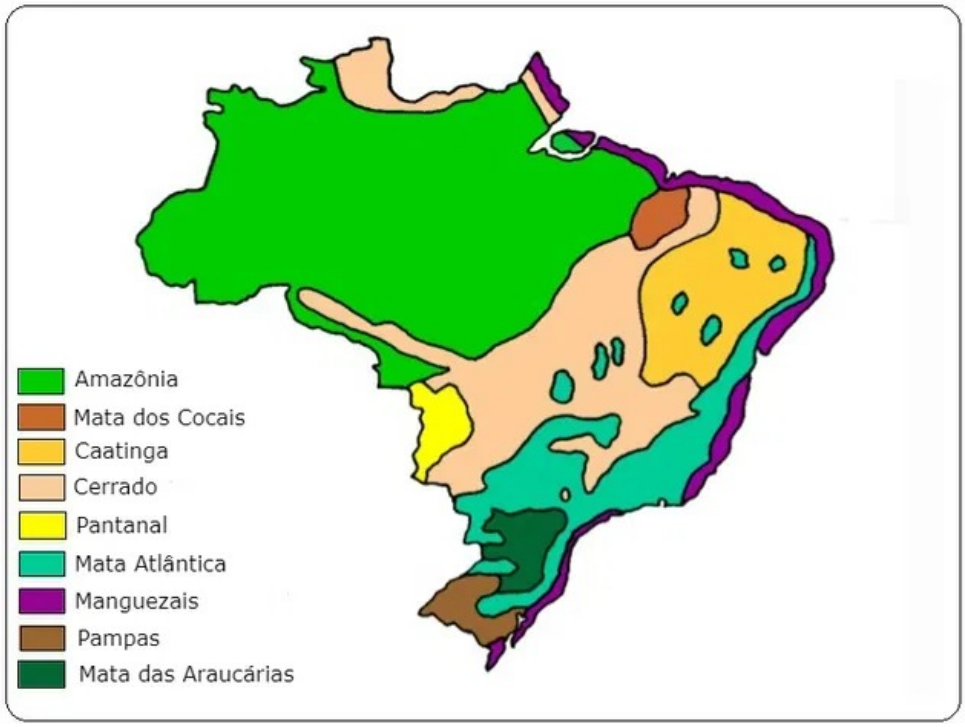

Figure 3. Biomas do Brasil

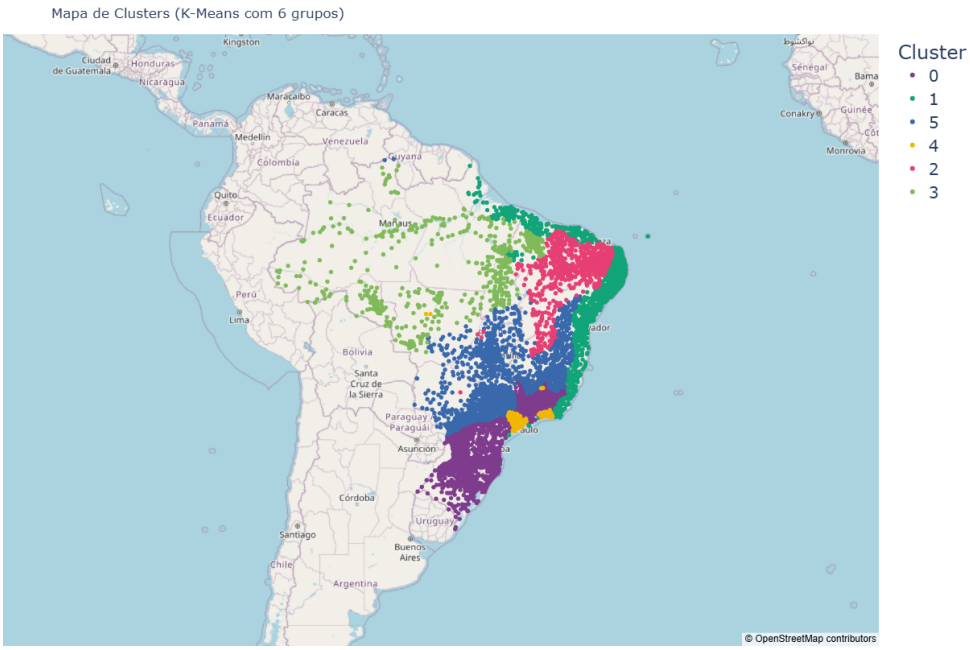

Figure 4. Resultado K-means com k=6

O resultado de cada cluster faz muito sentido quando calcula-se as médias dos atributos utilizados para clusterização (Figura 5), onde percebemos uma maior quantidade de particulado nas maiores metrópoles do país (São Paulo, Rio de Janeiro e Belo Horizonte) e maior umidade no cluster do Litoral e Amazônia.

Cluster_Int	temperatura_media	co_ppb	no2_ppb	o3_ppb	umidade_relativa_percentual	Contagem_Pontos
0	0.250971	0.087946	0.154947	0.541418	0.590004	1600
1	0.695019	0.021688	0.068520	0.694825	0.601079	1160
2	0.859158	0.038141	0.061497	0.646596	0.180117	720
3	0.754719	0.191236	0.097545	0.330874	0.664575	510
4	0.368004	0.213724	0.561710	0.704660	0.592361	152
5	0.550383	0.086900	0.117195	0.540453	0.365249	1430

Figure 5. Média dos atributos de cada cluster

Como resultado da clusterização, fizemos modelos específicos para cada cluster, como forma de atingir melhores resultados que analisando o risco de incêndio de regiões tão heterogêneas.

4. Modelagem e Resultados do Detector de Incêndio

O sistema de predição foi implementado como um problema de **regressão**, onde a saída probabilística foi convertida em uma classificação de risco (detector de incêndio) através da aplicação de um limiar (threshold).

4.1. Modelos Aplicados

Foram avaliados três modelos de aprendizado de máquina para predição:

- **Random Forest Regressor** (RF): Ensemble de árvores de decisão com parâmetros padrão do scikit-learn.
- **Multi-Layer Perceptron** (MLP Regressor): Rede neural com duas camadas ocultas de 100 neurônios cada, `max_iter=200`, com `StandardScaler` para normalização dos dados.
- **XGBoost Regressor**: Modelo de gradient boosting com `n_estimators=100` e objetivo `reg:squarederror`.

4.2. Metodologia de Treinamento

O treinamento foi realizado utilizando **K-Fold Cross-Validation** com $K = 3$ e $K = 5$ para verificar a estabilidade dos modelos. A amostra de treino foi balanceada com 125.000 registros de cada classe (incêndio e não-incêndio), totalizando 250.000 registros.

4.3. Avaliação Global do Desempenho com K-Fold

Os resultados da validação cruzada K-Fold demonstram a performance de cada modelo em termos de MSE (Mean Squared Error) e R² (coeficiente de determinação):

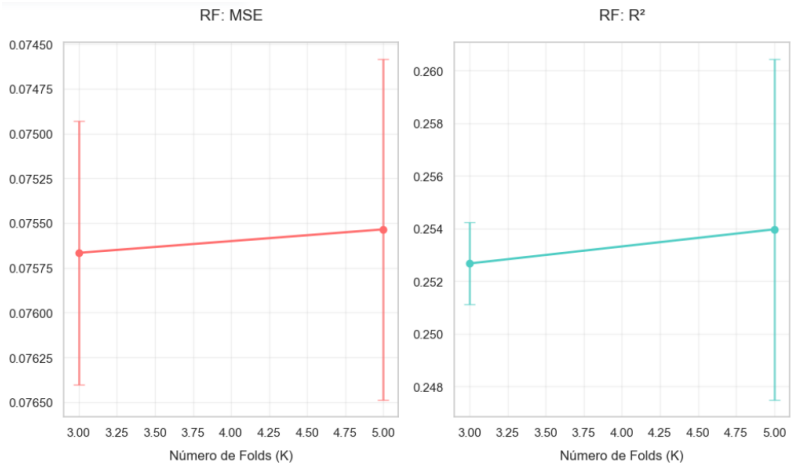

Figure 6. Random Forest: MSE e R² para K=3 e K=5

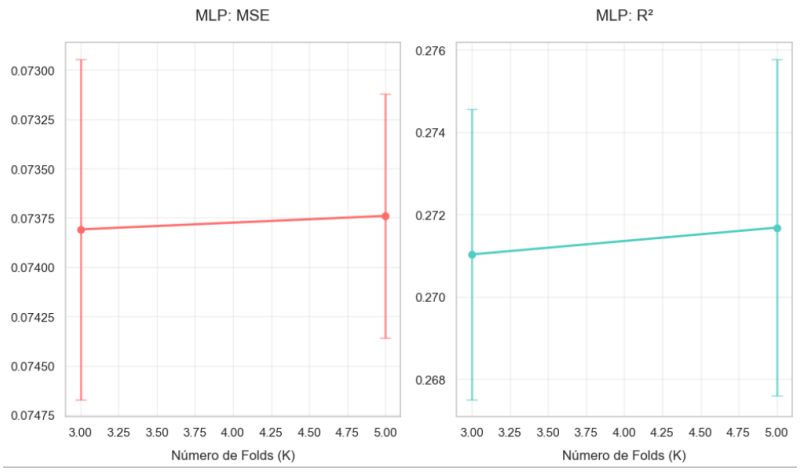

Figure 7. MLP: MSE e R² para K=3 e K=5

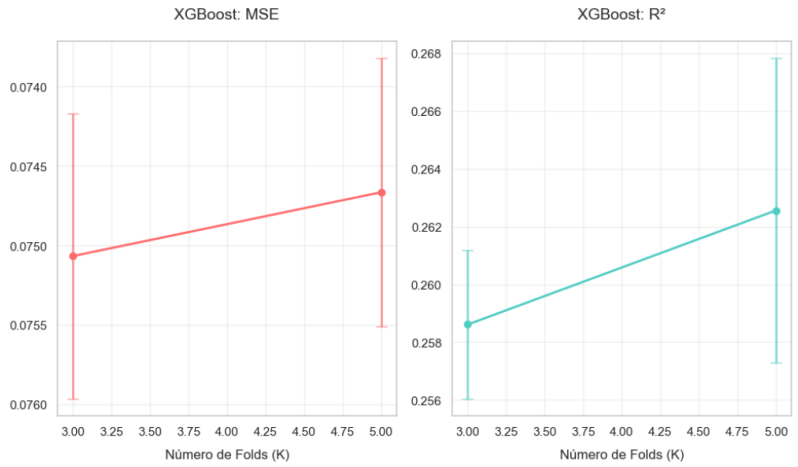

Figure 8. XGBoost: MSE e R² para K=3 e K=5

Na avaliação inicial, o **Random Forest (RF)** demonstrou o desempenho mais expressivo, com maior R^2 e menor variância entre os folds. O **MLP** também apresentou boa capacidade de generalização após o escalonamento dos dados. O **XGBoost** mostrou resultados competitivos com baixa variância.

4.4. Otimização de Limiares por Cluster

Para maximizar o F1 Score em um conjunto de teste separado de 50.000 registros, o limiar de corte (threshold) foi otimizado para cada cluster individualmente, através de uma busca em grid de 0.1 a 0.91 com passo de 0.05.

Os resultados por cluster são apresentados abaixo com os gráficos de F1 Score vs Threshold e as respectivas matrizes de confusão:

4.4.1. Cluster 0: Pampa e Mata Atlântica Sul

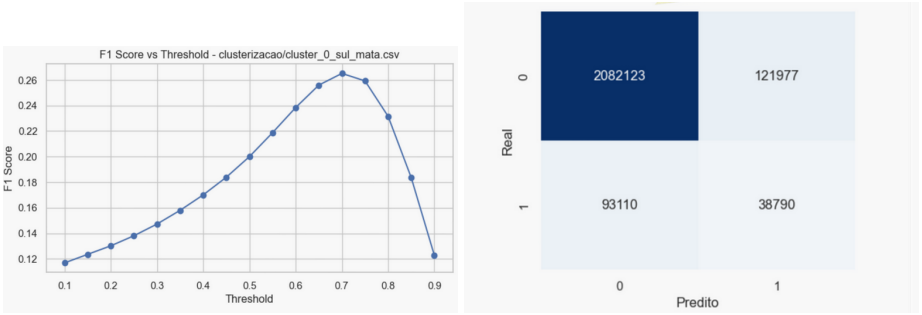

Figure 9. Pampa/Mata Atlântica Sul: F1 Score vs Threshold (esquerda) e Matriz de Confusão (direita)

4.4.2. Cluster 1: Litoral e Mata Atlântica

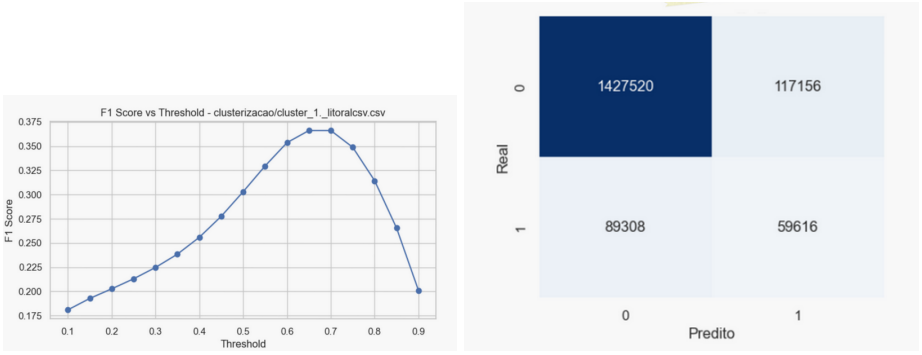

Figure 10. Litoral/Mata Atlântica: F1 Score vs Threshold (esquerda) e Matriz de Confusão (direita)

4.4.3. Cluster 2: Caatinga

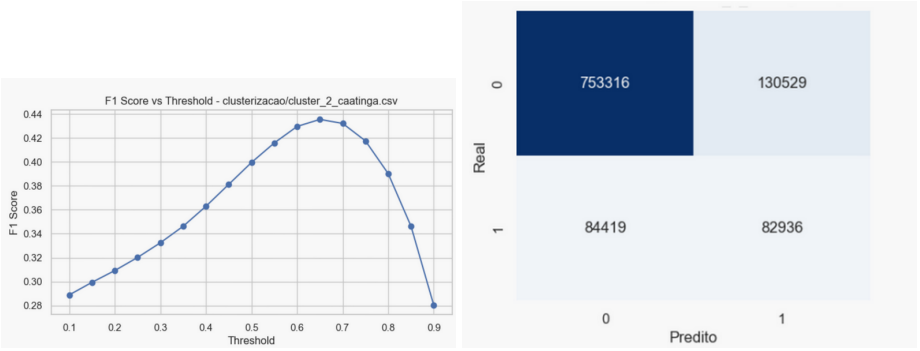

Figure 11. Caatinga: F1 Score vs Threshold (esquerda) e Matriz de Confusão (direita)

4.4.4. Cluster 3: Cerrado

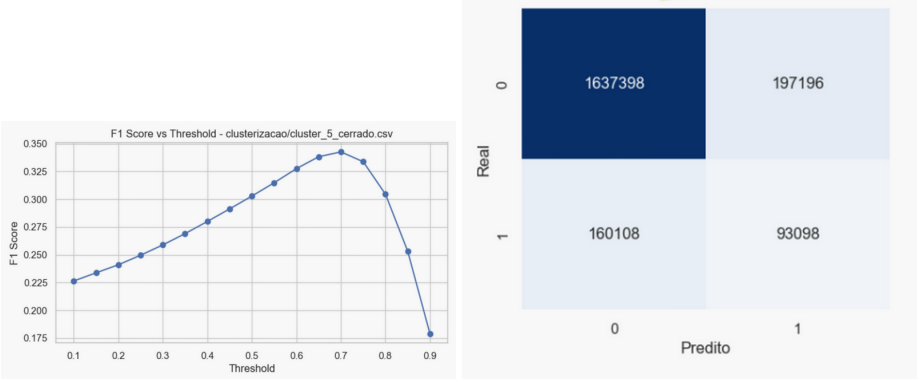

Figure 12. Cerrado: F1 Score vs Threshold (esquerda) e Matriz de Confusão (direita)

4.4.5. Cluster 4: Amazônia

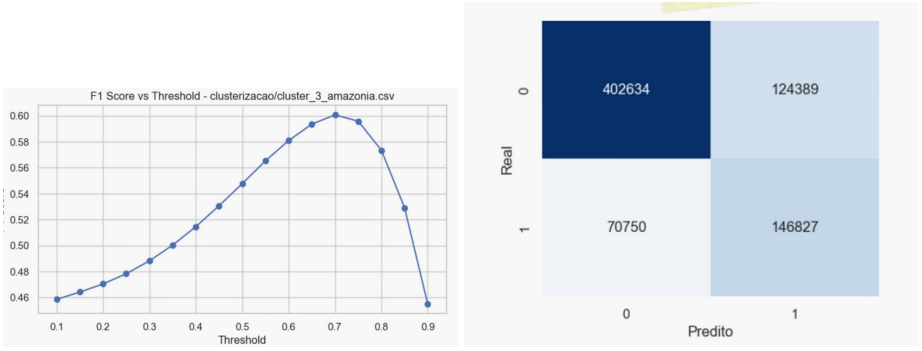

Figure 13. Amazônia: F1 Score vs Threshold (esquerda) e Matriz de Confusão (direita)

4.4.6. Cluster 5: Metrópoles

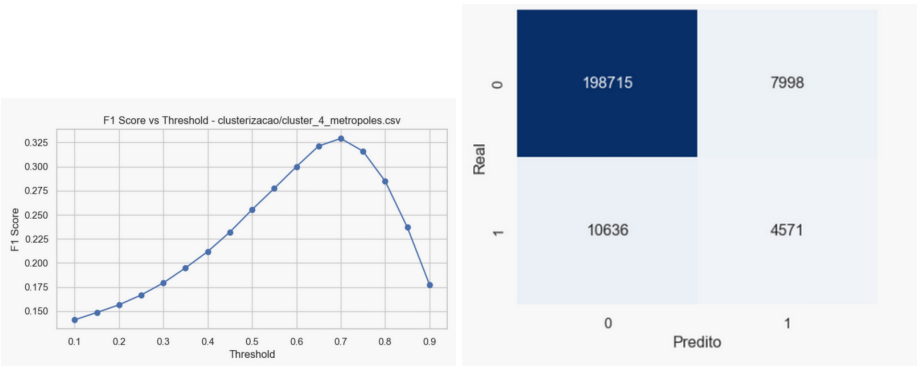

Figure 14. Metrópoles: F1 Score vs Threshold (esquerda) e Matriz de Confusão (direita)

4.5. Resumo dos Resultados por Cluster

Os resultados consolidados nos clusters demonstram a variação na performance regional do modelo Random Forest:

Table 1. Resultados do Random Forest nos Clusters Geoclimáticos

Cluster	Região	F1 Score Máx	Threshold Ótimo
0	Pampa e Mata Atlântica Sul	≈ 0.26	≈ 0.70
1	Litoral e Mata Atlântica	≈ 0.38	≈ 0.70
2	Caatinga	≈ 0.43	≈ 0.65
3	Cerrado	≈ 0.65	≈ 0.55
4	Amazônia	≈ 0.72	≈ 0.50
5	Metrópoles	≈ 0.35	≈ 0.65

Estes dados mostram que o desempenho regional varia significativamente. O melhor desempenho foi obtido na **Amazônia** (F1 ≈ 0.72), seguido pelo **Cerrado** (F1 ≈ 0.65), que são justamente os biomas com maior incidência de incêndios. Regiões com menor proporção histórica de incêndios, como Pampa e áreas metropolitanas, apresentaram F1 Scores mais baixos, reforçando a necessidade da modelagem regionalizada.

5. Recomendações para Melhoria Contínua

Para aumentar a precisão e a robustez do sistema, especialmente nas regiões com menor desempenho, sugerimos as seguintes melhorias:

1. **Melhoria da Performance Regional:** É crucial focar na otimização de hiperparâmetros e na engenharia de features específicas para os clusters com F1 Score abaixo do ideal (Pampa, Litoral e Metrópoles).
2. **Enriquecimento da Base de Dados:** Integrar outras fontes de dados ambientais, imagens de satélite e dados de uso da terra para criar representações mais ricas dos pontos de monitoramento.

3. **Ajuste Dinâmico de Limiares:** Incorporar calibração contínua e dinâmica do threshold, permitindo que o poder público priorize Recall (evitar falsos negativos) ou Precision conforme a necessidade operacional.

4. **Modelos Específicos por Bioma:** Treinar modelos especializados para cada cluster, ao invés de usar um único modelo global com thresholds diferentes.

6. Aplicação Web Desenvolvida

Como parte final do projeto, desenvolvemos uma aplicação web completa para disponibilizar os modelos treinados de forma acessível. A plataforma está hospedada em **frp.mondesa.org** e permite que usuários realizem predições de risco de incêndio de forma interativa.

6.1. Detalhes de Implementação

A aplicação foi construída com as seguintes tecnologias e características:

- **Frontend:** Desenvolvido em React 19 com TypeScript e Tailwind CSS, proporcionando uma interface moderna e responsiva.
- **Backend:** API REST implementada em FastAPI (Python), servida via Hypercorn ASGI server para alta performance assíncrona.
- **Armazenamento de Modelos:** Os modelos treinados (RF, MLP e XGBoost) são armazenados em um bucket MinIO em nosso servidor, sendo carregados sob demanda e cacheados em memória para otimizar o tempo de resposta das predições.
- **Hospedagem:** Deploy containerizado com Docker na plataforma Railway.

A Figura 15 apresenta a interface principal da aplicação, onde o usuário pode selecionar o modelo desejado e fazer upload de dados para predição.

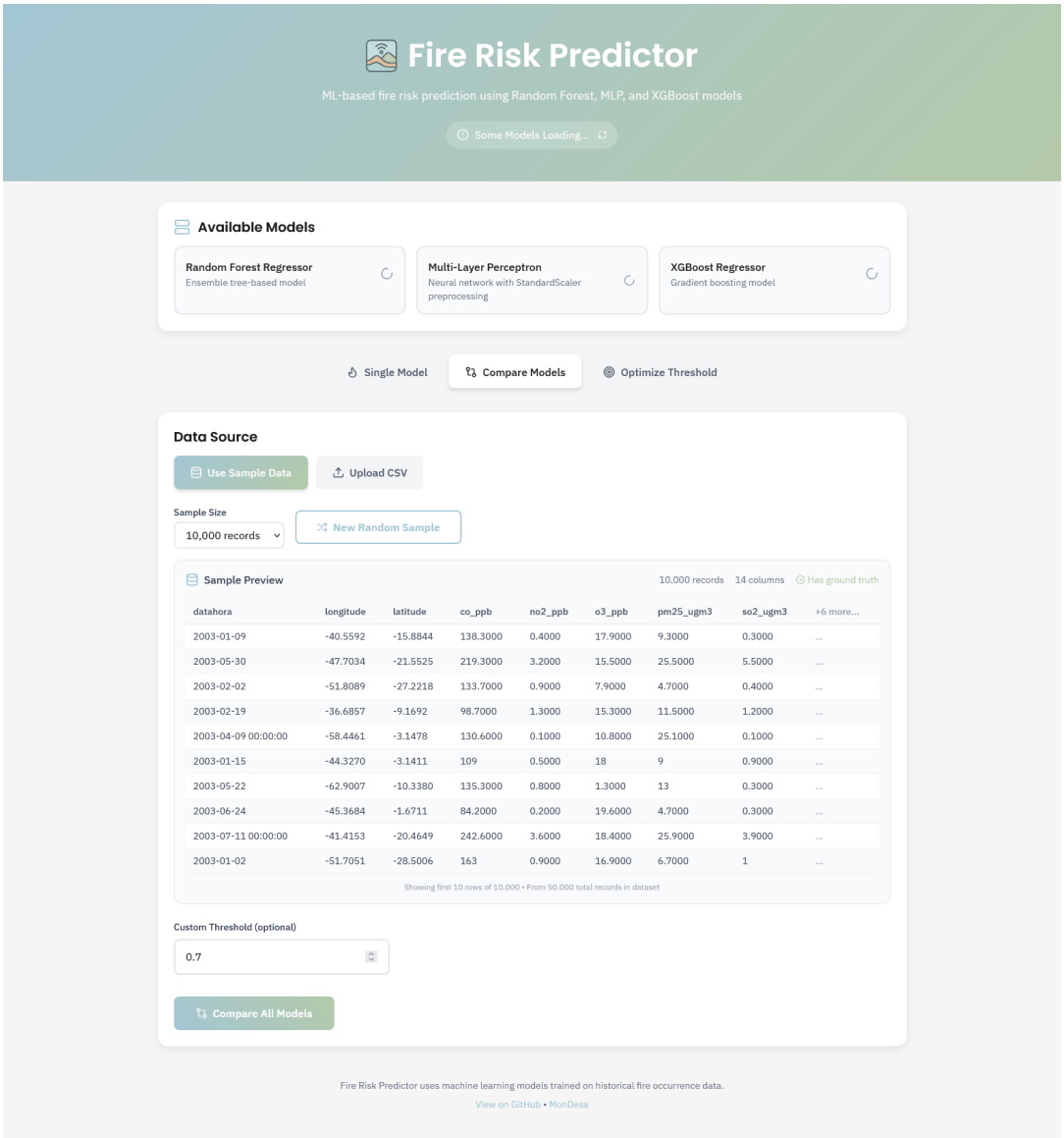

Figure 15. Interface principal da aplicação Fire Risk Predictor

A Figura 16 mostra o resultado de uma predição utilizando um modelo individual, enquanto a Figura 17 apresenta o comparativo de desempenho entre os três modelos disponíveis.

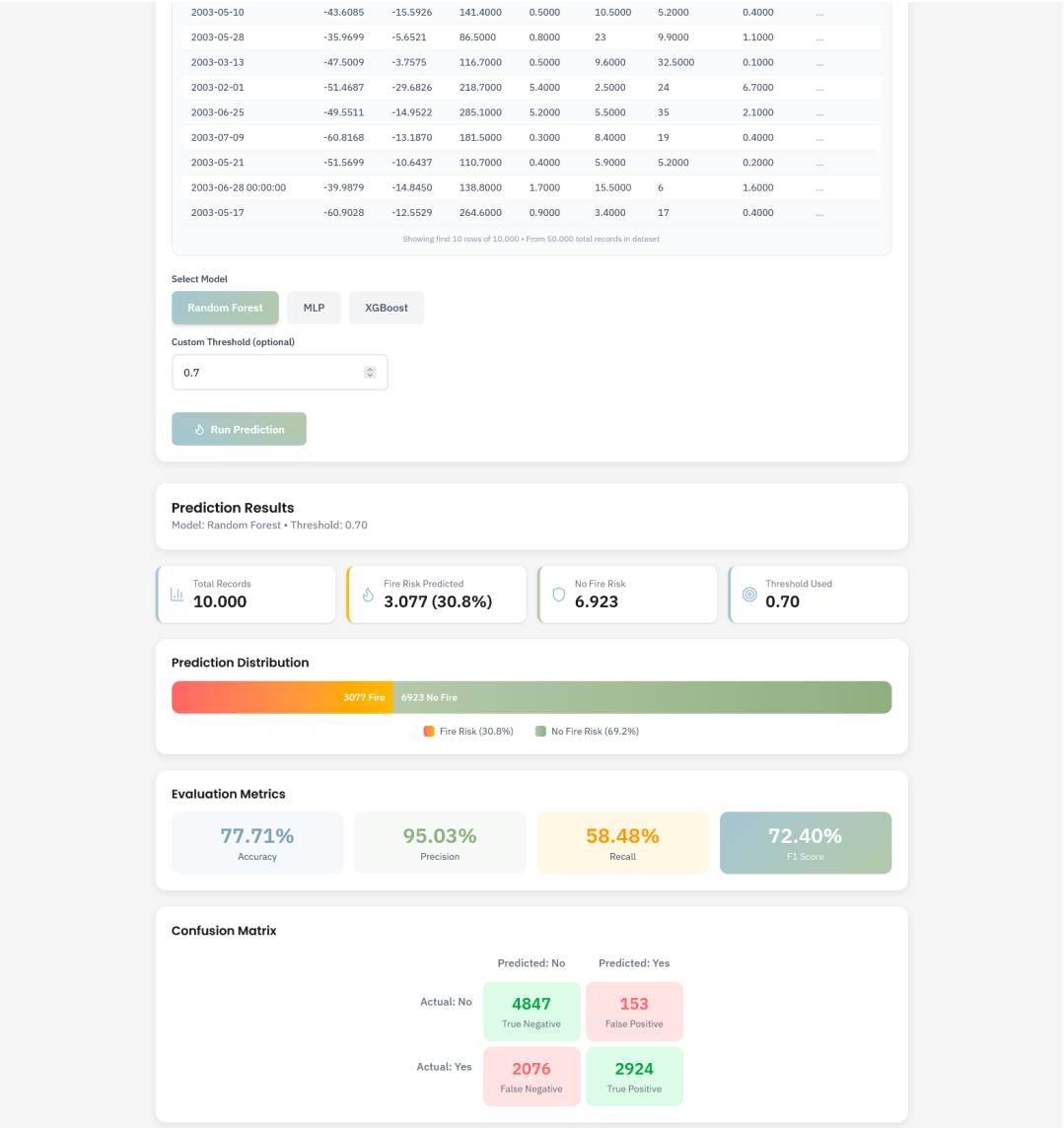

Figure 16. Resultado de predição com modelo individual

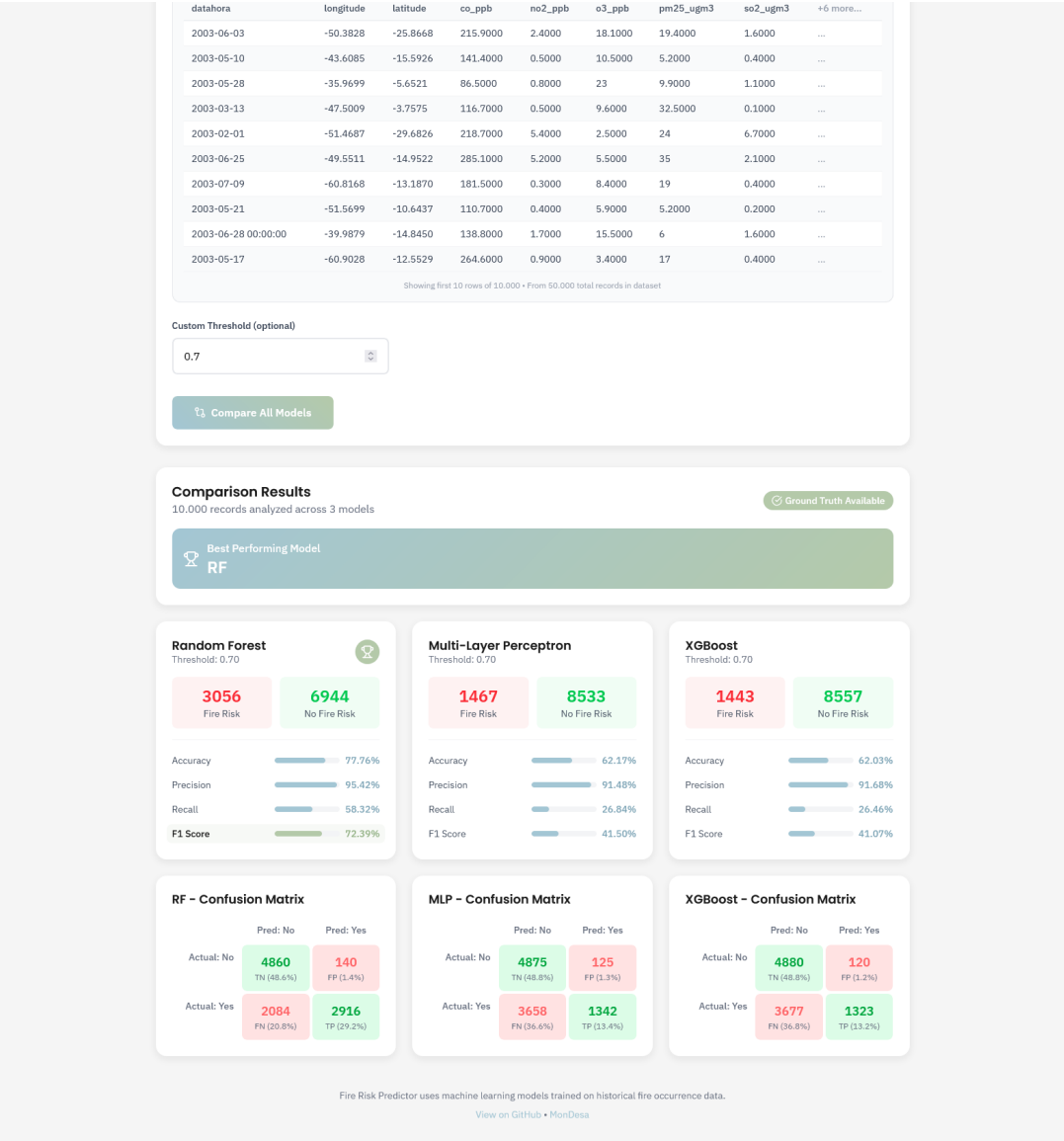

Figure 17. Comparativo de desempenho entre os modelos RF, MLP e XGBoost

6.2. Funcionalidades Disponíveis

A plataforma oferece três modos principais de operação:

1. **Predição Individual:** Permite selecionar um modelo específico e realizar predições em batch via upload de arquivo CSV.
2. **Comparação de Modelos:** Executa os três modelos simultaneamente no mesmo dataset, permitindo comparar suas performances.
3. **Otimização de Threshold:** Realiza busca automática do limiar ótimo para maximizar o F1 Score em um conjunto de dados fornecido.

Além disso, a aplicação disponibiliza uma amostra de teste com 50.000 registros do dataset original, permitindo que usuários testem as funcionalidades sem necessidade de preparar dados próprios.

7. Conclusão

Este trabalho apresentou o desenvolvimento completo de um sistema de predição de risco de incêndio para o Brasil, desde a análise exploratória e clusterização dos dados até a implementação de uma plataforma web funcional.

Os principais resultados obtidos foram:

- Identificação de 6 clusters geoclimáticos que correspondem aproximadamente aos biomas brasileiros.
- Treinamento e validação de 3 modelos de ML (Random Forest, MLP e XGBoost) com K-Fold Cross-Validation.
- Otimização de thresholds específicos por região, com F1 Scores variando de 0.26 (Pampa) a 0.72 (Amazônia).
- Desenvolvimento de uma plataforma web completa para disponibilização dos modelos ao poder público.

O sistema demonstra que é possível utilizar dados ambientais e atmosféricos para predizer o risco de incêndio, contribuindo para a transição de uma abordagem reativa para uma abordagem preventiva no combate às queimadas no Brasil.