INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DO PIAUÍ
CAMPUS FLORIANO
CURSO ANÁLISE E DESENVOLVIMENTO DE SISTEMAS

DIOGO BRUNO DE SÁ AMORIM

APLICAÇÃO DE MACHINE LEARNING EM DADOS CLIMÁTICOS PARA A PREDIÇÃO DE INCÊNDIOS NA CAATINGA PIAUIENSE.

FLORIANO
2026

DIOGO BRUNO DE SÁ AMORIM

APLICAÇÃO DE MACHINE LEARNING EM DADOS CLIMÁTICOS PARA A PREDIÇÃO DE INCÊNDIOS NA CAATINGA PIAUIENSE

Trabalho de Conclusão de Curso (artigo científico) apresentado como exigência parcial para obtenção do diploma do Curso de Análise e Desenvolvimento de Sistemas do Instituto Federal de Educação, Ciência e Tecnologia do Piauí, Campus Floriano .

Orientador: Prof. Dr. Justino Duarte Santos.

FLORIANO

2026

APLICAÇÃO DE MACHINE LEARNING EM DADOS CLIMÁTICOS PARA A PREDIÇÃO DE INCÊNDIOS NA CAATINGA PIAUIENSE

Diogo Bruno de Sá Amorim[1]
Justino Duarte Santos[2]

RESUMO

Os incêndios florestais representam uma importante ameaça ao bioma Caatinga, causando impactos ambientais, sociais e econômicos. Neste contexto, este trabalho teve como objetivo desenvolver um modelo preditivo baseado em aprendizado de máquina para estimar a ocorrência diária de incêndios florestais no estado do Piauí a partir de dados meteorológicos. Para isso, foram utilizados registros históricos de focos de calor e informações de estações meteorológicas, contemplando variáveis como temperatura, umidade relativa do ar, precipitação, pressão atmosférica e velocidade do vento. Após o pré-processamento e a integração dos dados, foi empregado o algoritmo *Light Gradient Boosting Machine*, com posterior otimização do limiar de decisão para melhorar o desempenho do modelo. Os resultados obtidos indicaram acurácia de 82%, precisão de 79%, sensibilidade de 88%, índice F1 de 83%, área sob a curva ROC de 90% e coeficiente de concordância kappa de 64%, demonstrando boa capacidade de identificação de eventos de incêndio. Como aplicação prática, o modelo foi utilizado na geração de mapas de risco, evidenciando seu potencial como ferramenta de apoio ao monitoramento ambiental e à tomada de decisão na prevenção e combate a incêndios florestais.
Palavras-chave: Aprendizado de Máquina; *Light Gradient Boosting Machine*; Incêndios Florestais; Caatinga; Predição de Risco.

ABSTRACT

Wildfires are a major threat to the Caatinga biome, causing significant environmental, social, and economic impacts. In this context, this study aimed to develop a machine learning-based predictive model to estimate the daily occurrence of wildfires in the state of Piauí using meteorological data. Historical wildfire records and weather station data were used, including variables such as air temperature, relative humidity, precipitation, atmospheric pressure, and wind speed. After data preprocessing and integration, the *Light Gradient Boosting* Machine algorithm was applied using training, validation, and testing datasets, together with decision threshold optimization to improve the model's predictive performance. The results showed an accuracy of 82%, precision of 79%, recall of 88%, an F1-score of 83%, an area under the curve of 90%, and a Cohen's kappa coefficient of 64%, demonstrating the model's ability to identify wildfire events. As a practical application, the model was used to generate wildfire risk maps, highlighting its potential as a decision-support tool for environmental monitoring and wildfire prevention.

[1] Graduando em Análise e Desenvolvimento de Sistemas. Instituto Federal de Educação, Ciência e Tecnologia do Piauí. diogobamorim06@gmail.com.
[2] Doutor. Professor do curso de Análise e Desenvolvimento de Sistemas do Instituto Federal de Educação, Ciência e Tecnologia do Piauí, Campus Floriano. justinoduarte@ifpi.edu.br

Keywords: Machine Learning; *Light Gradient Boosting Machine*; Wildfires; Caatinga; Risk Prediction.

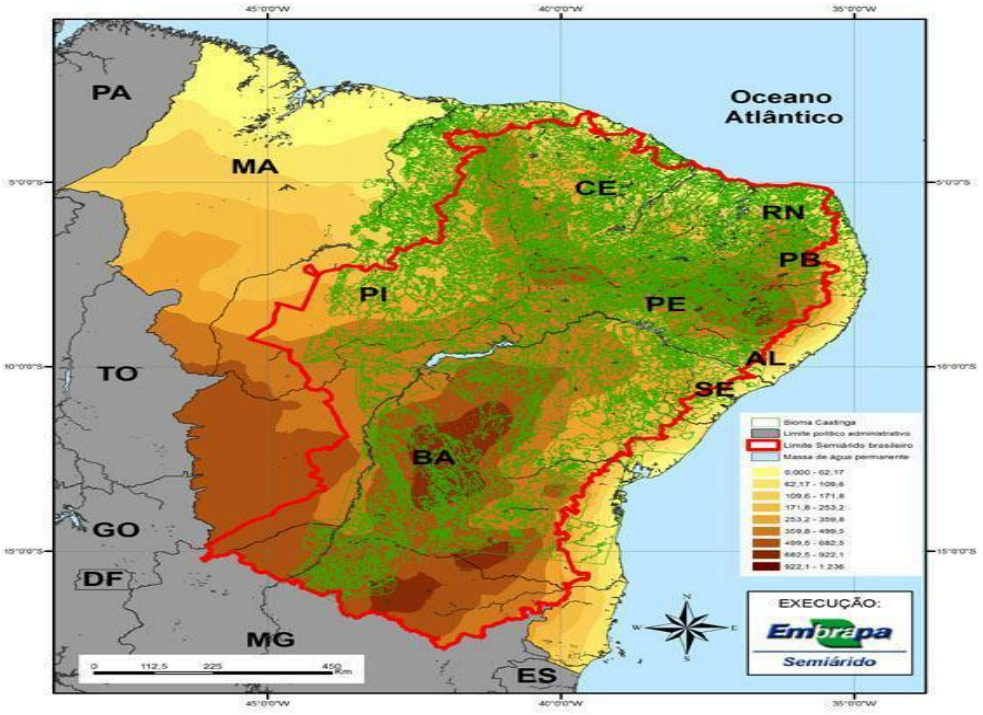
Data de aprovação: 09/07/2026

1 INTRODUÇÃO

A Caatinga é o único bioma exclusivamente brasileiro e ocupa uma parcela significativa da Região Nordeste, abrangendo extensa área do Estado do Piauí (INSA, 2025). As características climáticas desse bioma, marcadas por elevadas temperaturas, precipitações irregulares e longos períodos de estiagem, favorecem a ocorrência de incêndios florestais, os quais provocam impactos ambientais, sociais e econômicos relevantes (EMBRAPA, 2021; ICMBIO, 2023).

A Figura 1 apresenta a delimitação geográfica do bioma Caatinga no território brasileiro, destacando sua distribuição espacial e evidenciando a abrangência da área de estudo considerada nesta pesquisa.

Figura 1 - Mapa da Caatinga Brasileira



Fonte: Embrapa (2021)

Nos últimos anos, tem-se observado um aumento na ocorrência de queimadas e incêndios florestais em áreas da Caatinga. Segundo dados do Banco de Dados de Queimadas (BDQueimadas), do Instituto Nacional de Pesquisas Espaciais (INPE), o Piauí concentra a maior proporção de focos de calor entre os estados inseridos no bioma, seguido pelo Ceará e pela Bahia. Entre 2019 e 2023, esses três estados responderam por aproximadamente 80% dos focos de calor registrados na Caatinga, evidenciando a elevada suscetibilidade da região, especialmente do Piauí, à ocorrência de incêndios florestais. Esse aumento reforça a necessidade de fortalecer estratégias de monitoramento, prevenção e resposta a esses fenômenos. Nesse contexto, a disponibilidade de informações meteorológicas históricas e de registros de focos de calor constitui uma importante fonte de dados para o desenvolvimento de modelos capazes de identificar condições favoráveis à ocorrência de incêndios florestais.

A utilização de técnicas de Inteligência Artificial (IA), especialmente de Aprendizado de Máquina *(Machine Learning – ML)*, tem se consolidado como uma alternativa promissora para a previsão de incêndios florestais (Hanamoto, 2025; Moura, 2025). Esses modelos são capazes de identificar padrões complexos em grandes volumes de dados e estimar a probabilidade de ocorrência de novos focos de calor, contribuindo para sistemas de alerta precoce e para o planejamento de ações de prevenção e combate aos incêndios.

Estudos recentes têm demonstrado resultados promissores na aplicação de técnicas de aprendizado de máquina para a previsão de incêndios florestais. Hanamoto(2025) obtiveram resultados superiores a 80% de acurácia utilizando algoritmos como *Random Forest*, *Support Vector Machine (SVM)* e *Gradient Boosting* para prever focos de incêndio a partir de variáveis meteorológicas. Duarte (2022) propôs um modelo baseado na integração entre aprendizado de máquina e lógica fuzzy para representar as incertezas inerentes aos fenômenos climáticos. Neira et al. (2024), por sua vez, empregaram redes neurais profundas e imagens de satélite para realizar a detecção automática de focos de calor em tempo real, evidenciando o potencial da Inteligência Artificial aplicada ao monitoramento ambiental.

Apesar dos avanços observados na literatura, ainda existem desafios importantes para a aplicação dessas metodologias na realidade da Caatinga. Grande parte dos modelos disponíveis foi desenvolvida considerando bases de dados internacionais ou regiões com características climáticas distintas do semiárido brasileiro. Além disso, muitos trabalhos concentram seus esforços na escolha dos algoritmos de classificação, dedicando menor atenção às etapas de preparação dos dados, balanceamento das classes e redução de vieses espaciais decorrentes da distribuição das estações meteorológicas e dos registros de focos de calor. Essas limitações podem comprometer a capacidade de generalização dos modelos quando aplicados em cenários regionais.

Dessa forma, observa-se uma lacuna quanto à disponibilidade de modelos desenvolvidos especificamente para as características ambientais da Caatinga piauiense. O desenvolvimento de metodologias que integrem dados meteorológicos históricos, estratégias adequadas de preparação das amostras e algoritmos de elevado desempenho representa uma oportunidade para aprimorar sistemas de previsão de incêndios florestais e fornecer informações mais confiáveis aos órgãos responsáveis pela prevenção e combate ao fogo.

Nesse sentido, este trabalho tem como objetivo desenvolver e validar um modelo preditivo para estimar o risco diário de ocorrência de incêndios florestais no Estado do Piauí utilizando dados meteorológicos históricos e registros de focos de calor. O modelo é baseado no algoritmo *Light Gradient Boosting Machine (LightGBM)*, escolhido por sua elevada capacidade de modelar relações não lineares, eficiência computacional e desempenho em problemas de classificação supervisionada. Como diferencial metodológico, propõe-se uma estratégia de preparação dos dados voltada à redução dos vieses espaciais associados à distribuição das estações meteorológicas e à geração das amostras negativas, buscando aumentar a capacidade de generalização do modelo e fornecer uma ferramenta de apoio às ações de monitoramento e prevenção de incêndios florestais na Caatinga.

2 REFERENCIAL TEÓRICO

2.1 O BIOMA CAATINGA E OS INCÊNDIOS FLORESTAIS

A Caatinga constitui o único bioma exclusivamente brasileiro, ocupando cerca de 10,1% do território nacional e aproximadamente 53% da Região Nordeste. Sua distribuição abrange os estados de Alagoas, Bahia, Ceará, Maranhão, Minas Gerais, Paraíba, Pernambuco, Piauí, Rio Grande do Norte e Sergipe, desempenhando importante papel ecológico devido à elevada biodiversidade e ao expressivo número de espécies endêmicas adaptadas às condições do clima semiárido (EMBRAPA, 2021; INSA, 2025).

O clima da Caatinga caracteriza-se por elevadas temperaturas ao longo do ano, precipitações irregulares e longos períodos de estiagem. Em grande parte da região, as chuvas concentram-se em poucos meses, enquanto o restante do ano apresenta déficit hídrico acentuado. Essas condições favorecem o ressecamento da vegetação, aumentando a disponibilidade de material combustível e a suscetibilidade à ocorrência e à propagação de incêndios florestais (EMBRAPA, 2021).

A vegetação da Caatinga apresenta adaptações fisiológicas que permitem sua sobrevivência em ambientes de baixa disponibilidade hídrica, incluindo perda sazonal das folhas, sistemas radiculares profundos e mecanismos de armazenamento de água. Embora essas características aumentem a resistência das espécies aos períodos de seca, a estiagem prolongada favorece o acúmulo de biomassa seca, formando material combustível altamente suscetível à ignição quando exposto a fontes naturais ou antrópicas (EMBRAPA, 2021).

Além das condições naturais do bioma, a ação humana representa um dos principais fatores associados à ocorrência de incêndios florestais. Entre as práticas mais frequentes destacam-se a utilização do fogo para limpeza de áreas agrícolas, renovação de pastagens, desmatamento e queimadas irregulares. Em condições de baixa umidade relativa do ar e temperaturas elevadas, essas atividades podem perder o controle e originar incêndios de grandes proporções, ampliando os impactos ambientais e socioeconômicos (ICMBIO, 2023).

Os incêndios florestais provocam alterações significativas nos ecossistemas da Caatinga. Entre os principais impactos destacam-se a redução da cobertura vegetal, a intensificação dos processos erosivos, a alteração das propriedades físicas e químicas do solo, a perda de habitats naturais e a diminuição da biodiversidade (ICMBIO, 2023).

Além disso, a emissão de gases de efeito estufa e de material particulado contribui para a degradação da qualidade do ar e para o agravamento das mudanças climáticas, ampliando os impactos ambientais decorrentes desses eventos (ICMBIO, 2023).

No Estado do Piauí, esses impactos assumem importância estratégica devido à presença de importantes unidades de conservação inseridas no bioma, como o Parque Nacional da Serra das Confusões e o Parque Nacional da Serra da Capivara. Essas áreas desempenham papel fundamental na conservação da biodiversidade da Caatinga e demandam ações permanentes de monitoramento, prevenção e combate aos incêndios florestais. Nesse contexto, o Instituto Chico Mendes de Conservação da Biodiversidade destaca o Plano de Manejo Integrado do Fogo (PMIF) como uma importante estratégia para reduzir os impactos ambientais e fortalecer as ações de prevenção e resposta aos incêndios (ICMBIO, 2023).

Dados do Banco de Dados de Queimadas do Instituto Nacional de Pesquisas Espaciais (BDQueimadas/INPE, 2024) evidenciam aumento no número de focos de calor registrados na Região Nordeste, principalmente durante os períodos de estiagem, indicando a forte influência das condições meteorológicas sobre a ocorrência desses eventos. De forma complementar, informações do projeto MapBiomas (2024) apontam crescimento da área queimada no Brasil nos últimos anos, reforçando a necessidade de estratégias capazes de antecipar situações de risco e subsidiar ações de prevenção.

Diante desse cenário, torna-se evidente a necessidade de ferramentas que permitam compreender os fatores associados à ocorrência dos incêndios florestais e apoiar o planejamento de ações preventivas. Entre essas ferramentas destacam-se

os métodos baseados em Inteligência Artificial e Aprendizado de Máquina, cuja fundamentação teórica é apresentada na próxima seção.

2.2 INTELIGÊNCIA ARTIFICIAL E MACHINE LEARNING

A Inteligência Artificial (IA) corresponde a um ramo da Ciência da Computação dedicado ao desenvolvimento de sistemas capazes de executar tarefas que normalmente requerem inteligência humana, como reconhecimento de padrões, classificação de informações, tomada de decisão e realização de previsões. Nas últimas décadas, o avanço da capacidade computacional, aliado ao crescimento da disponibilidade de dados, impulsionou significativamente a aplicação dessas técnicas em diferentes áreas do conhecimento, incluindo saúde, agricultura, indústria, transporte e monitoramento ambiental (Moura, 2025).

Entre as principais áreas da Inteligência Artificial destaca-se o Aprendizado de Máquina (*Machine Learning – ML*), que reúne algoritmos capazes de aprender automaticamente padrões presentes em conjuntos de dados. Diferentemente da programação tradicional, na qual as regras de decisão são explicitamente definidas pelo desenvolvedor, os algoritmos de aprendizado de máquina ajustam seus parâmetros durante a fase de treinamento, permitindo construir modelos capazes de realizar previsões ou classificações a partir de novas observações (Dutra, 2024).

As técnicas de aprendizado de máquina podem ser classificadas em três categorias principais: aprendizado supervisionado, aprendizado não supervisionado e aprendizado por reforço. No aprendizado supervisionado, o algoritmo é treinado utilizando um conjunto de exemplos previamente rotulados, em que são conhecidas tanto as variáveis de entrada quanto a saída esperada. Durante o treinamento, o modelo identifica relações entre essas variáveis e aprende padrões capazes de estimar a classe de novos registros ainda não observados (Hanamoto, 2025). Essa abordagem é amplamente empregada em problemas de classificação e regressão, sendo uma das técnicas mais utilizadas em aplicações de aprendizado de máquina.

O desenvolvimento de modelos supervisionados envolve diversas etapas, iniciando pela obtenção e preparação dos dados. Nessa fase são realizadas atividades como integração de diferentes bases de dados, tratamento de valores

ausentes, construção de atributos derivados, organização temporal das observações e divisão do conjunto de dados em subconjuntos destinados ao treinamento, à validação e ao teste. Em seguida, o algoritmo é treinado utilizando o conjunto de treinamento, enquanto o conjunto de validação é empregado para monitorar o desempenho do modelo e ajustar seus parâmetros. Por fim, o conjunto de teste, mantido isolado durante todo o processo de desenvolvimento, é utilizado exclusivamente para avaliar a capacidade de generalização do modelo em dados não utilizados durante o treinamento.

Na área ambiental, técnicas de aprendizado de máquina têm sido amplamente empregadas na previsão de eventos extremos, no monitoramento de recursos naturais, na estimativa de produtividade agrícola e na previsão de secas, enchentes e incêndios florestais. Segundo Moura (2025), esses algoritmos apresentam elevada capacidade para processar grandes volumes de dados, identificar relações complexas entre variáveis climáticas e fornecer subsídios para sistemas de apoio à tomada de decisão no monitoramento ambiental.

No caso específico da previsão de incêndios florestais, diversos estudos têm demonstrado resultados promissores com a utilização de algoritmos supervisionados. Hanamoto(2025), por exemplo, obtiveram desempenho superior a 80% de acurácia na previsão de incêndios utilizando algoritmos baseados em árvores de decisão aplicados a dados meteorológicos. Esses estudos evidenciam o potencial da Inteligência Artificial como ferramenta de apoio ao desenvolvimento de sistemas de alerta precoce e ao planejamento de ações preventivas voltadas à gestão ambiental.

Entre os diferentes algoritmos utilizados em problemas de classificação supervisionada, destacam-se aqueles baseados na técnica de Gradient Boosting, que têm apresentado elevado desempenho em aplicações envolvendo dados tabulares. Esses algoritmos constroem sucessivas árvores de decisão de forma sequencial, buscando reduzir os erros de predição e melhorar o desempenho do modelo. Entre eles, destaca-se o *LightGBM*, reconhecido por sua eficiência computacional, escalabilidade e elevado desempenho em problemas de classificação e regressão envolvendo dados tabulares (Ke et al., 2017). Seus

princípios de funcionamento e suas principais características são apresentados na Seção 2.3.

2.3 GRADIENT BOOSTING E LIGHTGBM

Os algoritmos baseados em *Gradient Boosting* destacam-se entre as principais técnicas de aprendizado supervisionado para problemas de classificação e regressão, especialmente em aplicações envolvendo dados tabulares. Diferentemente de métodos baseados em uma única árvore de decisão, o *Gradient Boosting* constrói um conjunto de árvores de forma sequencial, em que cada nova árvore é treinada para corrigir os erros produzidos pelas árvores anteriormente geradas. Esse processo iterativo reduz gradativamente o erro de predição e melhora o desempenho do modelo, tornando-o capaz de representar relações complexas entre as variáveis de entrada e a variável de saída (Ke et al., 2017).

A principal característica dessa abordagem consiste na combinação de diversos modelos fracos (weak learners), geralmente representados por árvores de decisão rasas, formando um único modelo robusto. Em cada etapa do treinamento, o algoritmo calcula os resíduos produzidos pelo conjunto atual de árvores e ajusta uma nova árvore para minimizar esses erros por meio de técnicas de otimização baseadas no gradiente da função de perda. Como resultado, cada árvore adicionada busca complementar as anteriores, contribuindo para o aprimoramento progressivo do modelo (Ke et al., 2017).

Nos últimos anos, diversas implementações do algoritmo *Gradient Boosting* foram desenvolvidas com o objetivo de aumentar a eficiência computacional e o desempenho preditivo. Entre as principais destacam-se o *XGBoost*, o CatBoost e o *LightGBM*, sendo este último amplamente empregado em aplicações que envolvem grandes volumes de dados devido ao seu reduzido tempo de treinamento, baixo consumo de memória e elevada capacidade de processamento (Ke et al., 2017).

O *LightGBM* foi desenvolvido pela Microsoft como uma implementação otimizada do algoritmo *Gradient Boosting Decision Tree (GBDT)*. Sua principal característica é a utilização de estratégias capazes de reduzir significativamente o

tempo de treinamento e o consumo de memória, mantendo elevado desempenho preditivo. Entre essas estratégias destacam-se o crescimento das árvores baseado em folhas (leaf-wise growth) e a discretização dos atributos contínuos em histogramas, permitindo acelerar o processo de divisão dos nós e reduzir o custo computacional durante o treinamento (Ke et al., 2017).

Enquanto algoritmos tradicionais expandem as árvores de forma balanceada, realizando divisões simultâneas em todos os níveis, o *LightGBM* adota uma estratégia denominada *leaf-wise*. Nesse método, a cada iteração é expandida apenas a folha que apresenta maior potencial de redução da função de perda. Essa estratégia permite construir árvores mais eficientes, frequentemente alcançando maior desempenho preditivo com um número menor de divisões quando comparada aos métodos tradicionais de crescimento por níveis (level-wise) (Ke et al., 2017).

Outra característica importante do *LightGBM* é a utilização de histogramas para representar os atributos contínuos. Em vez de avaliar todos os valores possíveis durante a construção das árvores, o algoritmo agrupa os valores em intervalos discretos (bins), reduzindo significativamente a quantidade de cálculos necessários para encontrar os melhores pontos de divisão. Essa abordagem proporciona menor consumo de memória e maior velocidade de processamento, tornando o algoritmo especialmente adequado para conjuntos de dados extensos e aplicações que demandam elevada eficiência computacional (Ke et al., 2017).

Além da eficiência computacional, o *LightGBM* oferece recursos que favorecem seu desempenho em diferentes problemas de classificação e regressão, incluindo mecanismos para tratamento de bases de dados desbalanceadas, controle do sobreajuste (*overfitting*) e ajuste de hiperparâmetros voltados à otimização do processo de treinamento. Essas características contribuem para sua ampla utilização em aplicações nas áreas de ciência de dados, sensoriamento remoto, monitoramento ambiental, previsão climática e análise de séries temporais (Ke et al., 2017).

Estudos recentes têm demonstrado o potencial dos algoritmos baseados em *Gradient Boosting* na modelagem de fenômenos ambientais complexos.

Hanamoto(2025), por exemplo, obtiveram desempenho superior a 80% de acurácia na previsão de incêndios florestais utilizando algoritmos dessa família aplicados a variáveis meteorológicas. Esses resultados evidenciam a capacidade desses modelos em representar relações não lineares entre variáveis climáticas e a ocorrência de incêndios, fornecendo suporte ao desenvolvimento de sistemas de monitoramento e alerta precoce.

Em virtude de sua elevada eficiência computacional, escalabilidade e capacidade de modelar relações complexas entre variáveis, o *LightGBM* consolidou-se como uma das principais técnicas de aprendizado supervisionado para problemas de classificação baseados em dados tabulares. Essas características justificam sua ampla adoção em aplicações de previsão e monitoramento ambiental, incluindo estudos relacionados à ocorrência de incêndios florestais.

2.4 TRABALHOS RELACIONADOS

A aplicação de técnicas de Inteligência Artificial na previsão de incêndios florestais tem apresentado avanços significativos nos últimos anos, impulsionada pela crescente disponibilidade de dados meteorológicos, imagens de satélite e registros históricos de focos de calor. Diversos estudos têm investigado o emprego de algoritmos de aprendizado supervisionado e de outras abordagens computacionais para apoiar sistemas de monitoramento ambiental e emissão de alertas preventivos. Nesta seção são apresentados alguns dos principais trabalhos relacionados ao tema desta pesquisa, destacando os métodos empregados, os dados utilizados e os resultados obtidos.

Hanamoto(2025) propõem um sistema de baixo custo para emissão de alertas preventivos de incêndios florestais baseado em algoritmos de aprendizado supervisionado. O estudo utiliza dados meteorológicos provenientes de estações oficiais, incluindo temperatura do ar, umidade relativa, precipitação, pressão atmosférica, radiação solar e velocidade do vento, associados aos registros históricos de focos de incêndio obtidos por sensoriamento remoto. Foram avaliados diferentes algoritmos de classificação, destacando-se o *Random Forest* e a *Support*

Vector Machine (SVM). Entre os modelos analisados, o *Random Forest* apresentou o melhor desempenho, alcançando aproximadamente 81,4% de acurácia, 79,4% de precisão, 83,9% de sensibilidade, 81,6% de F1-Score e AUC de 88,7%, evidenciando o potencial das técnicas de aprendizado de máquina na previsão de incêndios florestais.

Duarte (2022) desenvolveu um modelo de suscetibilidade a incêndios florestais utilizando técnicas de Inteligência Artificial associadas à lógica fuzzy. O objetivo foi representar as incertezas inerentes aos fenômenos ambientais por meio de regras de inferência capazes de integrar diferentes variáveis meteorológicas. O modelo utilizou principalmente temperatura do ar, umidade relativa e precipitação como variáveis de entrada, demonstrando que esses atributos exercem influência significativa sobre a suscetibilidade ao fogo. Os resultados reforçam a importância da utilização de variáveis climáticas na modelagem preditiva de incêndios, embora o trabalho não apresente uma implementação voltada à construção de sistemas operacionais de previsão.

Neira et al. (2024) investigaram a utilização de técnicas de Deep Learning e visão computacional para a detecção automática de focos de incêndio a partir de imagens. O estudo empregou arquiteturas da família *YOLO (You Only Look Once)*, treinadas com conjuntos de imagens contendo fogo e fumaça. Os resultados demonstraram elevados valores de precisão, recall e F1-Score, indicando que modelos baseados em redes neurais convolucionais apresentam elevado desempenho na identificação automática de focos de incêndio em tempo real.

Moura (2025) discute o papel da Inteligência Artificial no monitoramento ambiental, destacando que algoritmos de aprendizado de máquina possuem capacidade para identificar padrões complexos em séries históricas de dados climáticos e fornecer suporte à tomada de decisão em sistemas ambientais. Segundo o autor, a aplicação dessas técnicas possibilita antecipar eventos extremos, como secas, enchentes e incêndios florestais, contribuindo para o desenvolvimento de sistemas de monitoramento e alerta precoce. O estudo ressalta ainda que modelos preditivos favorecem a otimização de recursos operacionais e o aumento da eficiência das estratégias de prevenção.

Além dos trabalhos voltados especificamente à previsão de incêndios florestais, estudos recentes evidenciam o elevado desempenho de algoritmos baseados em *Gradient Boosting* em problemas de classificação envolvendo dados tabulares. Ke et al. (2017) demonstraram que o *LightGBM* apresenta elevada eficiência computacional e alto desempenho preditivo devido à utilização de estratégias como crescimento por folhas (*leaf-wise growth*) e discretização dos atributos em histogramas. Essas características tornam o algoritmo adequado para aplicações que envolvem grandes volumes de dados e relações complexas entre variáveis.

O Quadro 1 apresenta uma síntese dos principais trabalhos analisados, destacando as técnicas empregadas, os tipos de dados utilizados, os resultados obtidos e as respectivas áreas de aplicação.

Quadro 1 - Síntese dos trabalhos relacionados

Trabalho	Técnicas Utilizadas	Tipo de Dados	Resultados/ Contribuições	Área de Aplicação
Hanamoto (2025)	Random Forest, SVM	Dados meteorológicos + focos de incêndio (sensoriamento remoto)	Acurácia ≈ 81,0%, AUC ≈ 88,7%; bom desempenho operacional para alerta preventivo	Áreas florestais do Brasil
Moura (2025)	Machine Learning (abordagem conceitual)	Séries históricas de dados climáticos	Identificação de padrões complexos e suporte à prevenção de desastres ambientais	Biomas Brasileiros
	Deep Learning	Imagens de satélite e	Alta precisão e recall na	Monitoramento ambiental

Neira et al. (2024)	(YOLO)	imagens rotuladas de fogo/fumaça	detecção automática de calor em tempo real	via satélite
Duarte (2022)	IA + Inferência Fuzzy	Dados meteorológicos	Modelagem eficiente da suscetibilidade ao fogo e validação das variáveis climáticas	Regiões suscetíveis a queimadas no Brasil
Ke *et al.* (2017)	*LightGBM*	Dados tabulares	Alta eficiência computacional e elevado desempenho em classificação	Diversas aplicações de Machine Learning

Fonte: Elaborado pelo autor (2026).

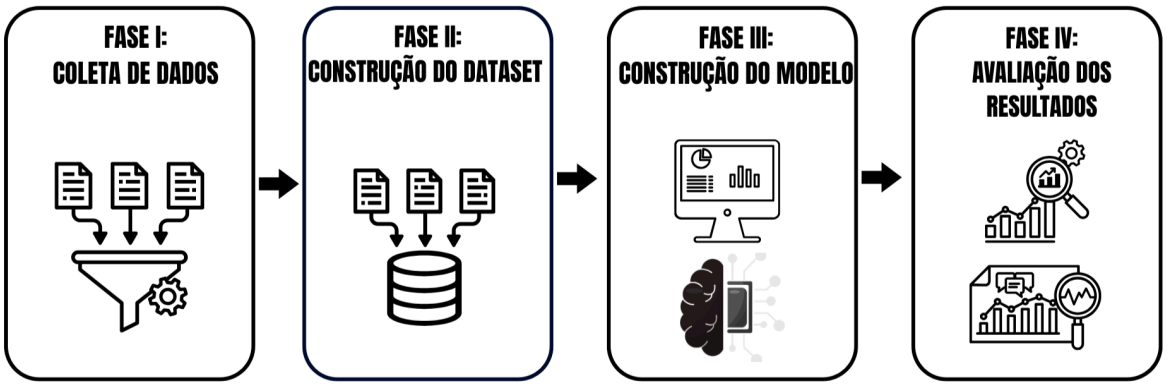
A análise dos trabalhos relacionados evidencia que técnicas de Inteligência Artificial vêm apresentando resultados consistentes na previsão e detecção de incêndios florestais. Entretanto, observa-se que a maior parte das pesquisas concentra-se na comparação entre algoritmos de classificação ou na utilização de imagens de satélite, havendo menor número de estudos direcionados às condições climáticas da Caatinga. Também são escassos os trabalhos que exploram estratégias para reduzir os vieses decorrentes da distribuição espacial das estações meteorológicas durante a construção do conjunto de dados.

Nesse contexto, a presente pesquisa diferencia-se por integrar dados meteorológicos históricos e registros de focos de incêndio por meio de uma estratégia de associação espacial entre cada foco e as três estações meteorológicas mais próximas. Além disso, emprega o algoritmo *LightGBM* para desenvolver um modelo preditivo direcionado às condições ambientais do estado do Piauí, buscando aumentar a representatividade espacial das variáveis utilizadas e contribuir para o desenvolvimento de ferramentas de apoio ao monitoramento e à prevenção de incêndios florestais.

3 PROCEDIMENTOS METODOLÓGICOS

O processo metodológico foi estruturado em fases interdependentes, conforme ilustrado na Figura 2, ele abrange: (I) A coleta e organização dos dados meteorológicos e de incêndios; (II) A construção do dataset utilizando dados de N estações meteorológicas; (III) Construção do modelo de Machine Learning; e (IV) A avaliação dos resultados por meio de métricas estatísticas.

Figura 2 - Processo metodológico



Fonte: Elaborado pelo autor (2026).

3.1 TIPO E ABORDAGEM DE PESQUISA

A presente pesquisa possui abordagem quantitativa, uma vez que se fundamenta na utilização de dados meteorológicos numéricos e registros históricos de incêndios para construção, treinamento e avaliação de modelos computacionais de classificação.

Quanto aos objetivos, caracteriza-se como pesquisa descritiva e exploratória, pois busca identificar padrões climáticos associados à ocorrência de incêndios florestais na Caatinga, além de investigar a aplicabilidade de algoritmos de aprendizado supervisionado para predição de risco.

Adicionalmente, trata-se de uma pesquisa aplicada, uma vez que propõe

solução computacional voltada ao monitoramento e apoio à prevenção de incêndios florestais. Também apresenta natureza documental, pois utiliza dados secundários provenientes de bases públicas institucionais.

As etapas de coleta, tratamento, integração, modelagem e experimentação foram conduzidas em ambiente computacional, utilizando ferramentas de programação e bibliotecas especializadas para análise de dados e aprendizado de máquina.

3.2 INSTRUMENTOS E PROCEDIMENTOS DE COLETA DE DADOS

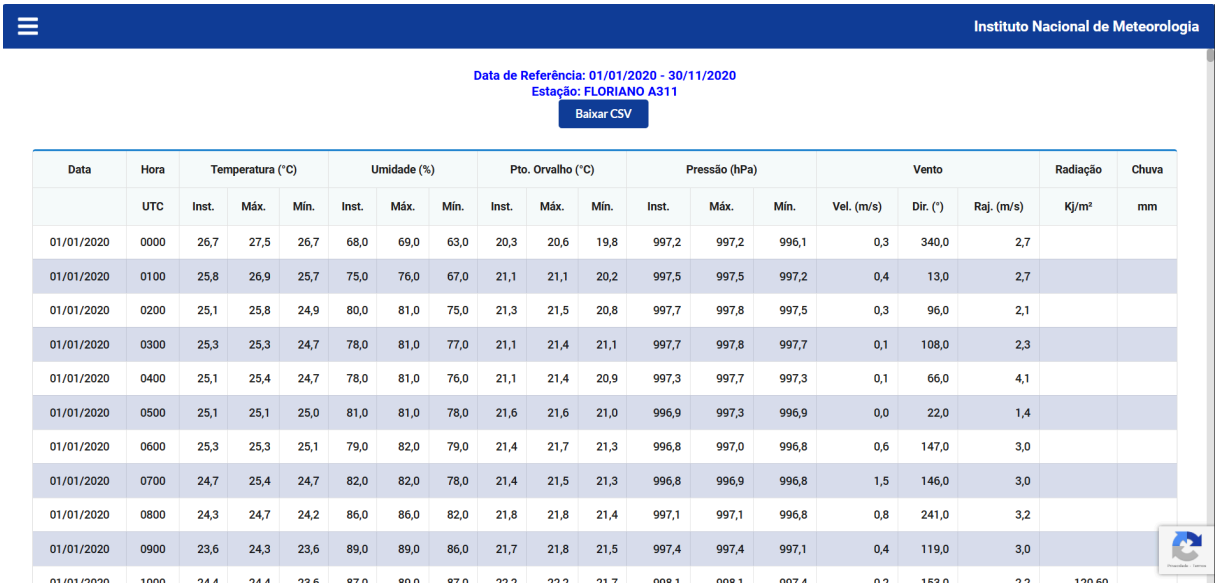
A coleta de dados envolveu duas fontes principais: registros meteorológicos diários e registros históricos oficiais de focos de incêndio, ambos disponibilizados publicamente por instituições governamentais brasileiras.

Os dados climáticos foram obtidos a partir das bases públicas do INMET, conforme ilustrado na Figura 3, contendo séries históricas diárias registradas de forma automatizada por estações meteorológicas distribuídas na região.

Já os dados referentes aos focos de incêndio foram obtidos por meio do Banco de Dados de Queimadas (BDQueimadas), disponibilizado pelo Instituto Nacional de Pesquisas Espaciais (INPE), conforme apresentado na Figura 4. Essa base reúne registros históricos de ocorrências de queimadas detectadas por satélites, permitindo acesso via plataforma web e download em formatos abertos.

A utilização conjunta dessas bases possibilitou integrar informações climáticas e registros de incêndio em uma única estrutura, permitindo construção de dataset supervisionado para treinamento e avaliação do modelo preditivo.

Figura 3 - Dados meteorológicos disponibilizados pelo inmet



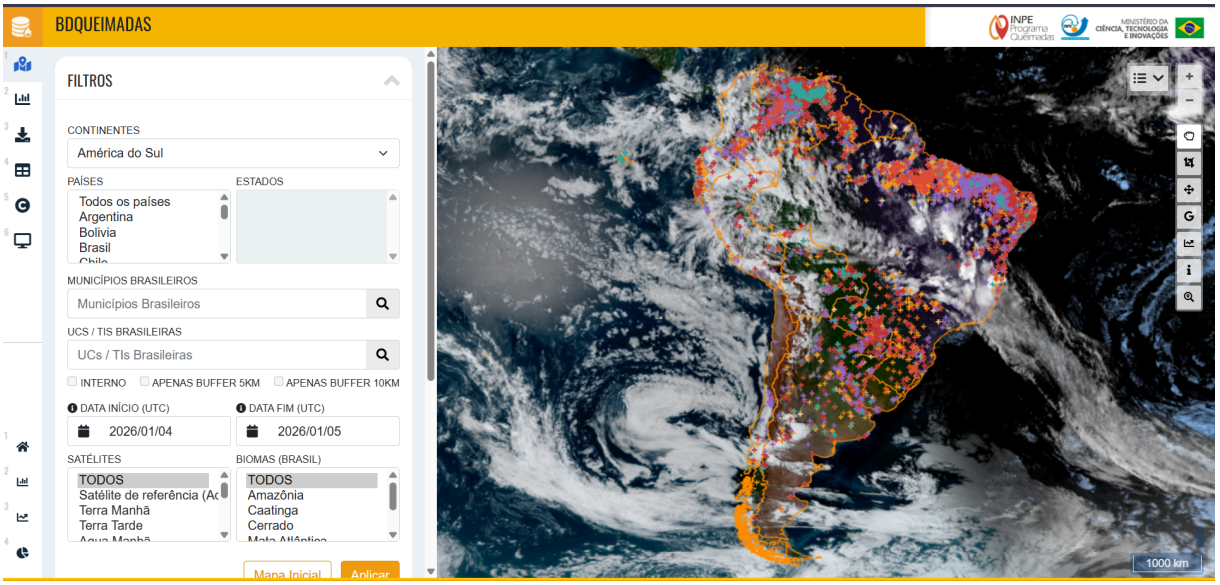
Instituto Nacional de Meteorologia

Data de Referência: 01/01/2020 - 30/11/2020
Estação: FLORIANO A311
Baixar CSV

Data	Hora	Temperatura (°C)			Umidade (%)			Pto. Orvalho (°C)			Pressão (hPa)			Vento			Radiação	Chuva
	UTC	Inst.	Máx.	Mín.	Inst.	Máx.	Mín.	Inst.	Máx.	Mín.	Inst.	Máx.	Mín.	Vel. (m/s)	Dir. (°)	Raj. (m/s)	Kj/m²	mm
01/01/2020	0000	26,7	27,5	26,7	68,0	69,0	63,0	20,3	20,6	19,8	997,2	997,2	996,1	0,3	340,0	2,7		
01/01/2020	0100	25,8	26,9	25,7	75,0	76,0	67,0	21,1	21,1	20,2	997,5	997,5	997,2	0,4	13,0	2,7		
01/01/2020	0200	25,1	25,8	24,9	80,0	81,0	75,0	21,3	21,5	20,8	997,7	997,8	997,5	0,3	96,0	2,1		
01/01/2020	0300	25,3	25,3	24,7	78,0	81,0	77,0	21,1	21,4	21,1	997,7	997,8	997,7	0,1	108,0	2,3		
01/01/2020	0400	25,1	25,4	24,7	78,0	81,0	76,0	21,1	21,4	20,9	997,3	997,7	997,3	0,1	66,0	4,1		
01/01/2020	0500	25,1	25,1	25,0	81,0	81,0	78,0	21,6	21,6	21,0	996,9	997,3	996,9	0,0	22,0	1,4		
01/01/2020	0600	25,3	25,3	25,1	79,0	82,0	79,0	21,4	21,7	21,3	996,8	997,0	996,8	0,6	147,0	3,0		
01/01/2020	0700	24,7	25,4	24,7	82,0	82,0	78,0	21,4	21,5	21,3	996,8	996,9	996,8	1,5	146,0	3,0		
01/01/2020	0800	24,3	24,7	24,2	86,0	86,0	82,0	21,8	21,8	21,4	997,1	997,1	996,8	0,8	241,0	3,2		
01/01/2020	0900	23,6	24,3	23,6	89,0	89,0	86,0	21,7	21,8	21,5	997,4	997,4	997,1	0,4	119,0	3,0		

Fonte: [3]Instituto Nacional de Meteorologia (INMET)

Figura 4 - Dados de focos de incêndios disponibilizados pelo INPE



Fonte - [4]Instituto Nacional de Pesquisa Espacial (INPE)

[3] Instituto Nacional de Meteorologia (INMET). Disponível em: INMET :: Tempo

[4] Instituto Nacional de Pesquisa Espacial (INPE). Disponível em: BDQueimadas - Programa Queimadas - INPE

3.2.1 Dados Meteorológicos

Os dados meteorológicos utilizados nesta pesquisa foram obtidos a partir das estações meteorológicas automáticas disponibilizadas pelo Instituto Nacional de Meteorologia (INMET), órgão oficial responsável pelo monitoramento das condições climáticas no território brasileiro. A escolha dessa base deve-se à sua confiabilidade institucional, à disponibilidade pública dos registros históricos e à abrangência espacial e temporal das informações disponibilizadas.

Para a construção do conjunto de dados foram utilizados registros diários compreendidos entre janeiro de 2015 e dezembro de 2019, período selecionado por apresentar disponibilidade simultânea de informações meteorológicas e registros de focos de incêndio necessários ao desenvolvimento do modelo preditivo.

Foram selecionadas variáveis meteorológicas diretamente relacionadas às condições ambientais favoráveis à ocorrência e à propagação de incêndios florestais, compreendendo precipitação total diária, pressão atmosférica média diária, temperatura do ponto de orvalho média diária, temperaturas máxima, média e mínima do ar, umidade relativa do ar média e mínima, rajada máxima diária do vento e velocidade média diária do vento.

A seleção dessas variáveis fundamenta-se em estudos que demonstram sua influência sobre a dinâmica dos incêndios florestais. A precipitação e a umidade relativa do ar estão associadas à disponibilidade hídrica da vegetação e ao grau de ressecamento do ambiente. As variáveis de temperatura influenciam diretamente a evaporação da água, o aquecimento da biomassa e a redução da umidade superficial. Por sua vez, a velocidade e a rajada do vento favorecem a propagação das chamas e o transporte de partículas incandescentes, ampliando a área potencialmente atingida pelo fogo. Esses fatores são amplamente reconhecidos como importantes condicionantes da ocorrência e da propagação de incêndios florestais (Duarte, 2022; Hanamoto, 2025).

Após a obtenção dos dados, os registros foram organizados em periodicidade diária, padronizando o formato temporal das observações e permitindo sua

integração com os registros históricos de focos de incêndio utilizados na construção do conjunto de dados.

3.2.2 Dados de Incêndios Florestais

Os registros de incêndios florestais foram obtidos junto ao BDQueimadas, disponibilizado pelo Instituto Nacional de Pesquisas Espaciais (INPE). Essa plataforma reúne registros georreferenciados de focos de calor detectados por satélites de sensoriamento remoto, constituindo uma das principais bases oficiais utilizadas para o monitoramento de queimadas e incêndios florestais no Brasil.

Assim como os dados meteorológicos, foram utilizados registros compreendidos entre janeiro de 2015 e dezembro de 2019, garantindo compatibilidade temporal entre as bases de dados empregadas na pesquisa.

Inicialmente, foram coletados os registros históricos de focos de incêndio referentes à área de estudo. Posteriormente, esses registros foram integrados aos dados meteorológicos por meio da correspondência espacial e temporal entre os focos de calor e as estações meteorológicas, possibilitando associar as condições climáticas observadas à ocorrência de cada evento.

Além da identificação das ocorrências de incêndio, a base também foi utilizada na construção das amostras sem ocorrência de fogo. Dessa forma, foi possível estruturar o problema como uma tarefa de classificação binária supervisionada, composta por registros representando tanto a presença quanto a ausência de incêndios florestais.

3.3 CONSTRUÇÃO E ORGANIZAÇÃO DO DATASET

A construção do dataset representa uma etapa central para o desempenho dos modelos preditivos, uma vez que a qualidade da base influencia diretamente a capacidade de aprendizado, generalização e identificação de padrões associados à ocorrência de incêndios florestais. Dessa forma, o processo de construção da base foi estruturado de modo a garantir coerência temporal, representatividade espacial e

consistência estatística, assegurando que as variáveis fornecidas aos algoritmos descrevam adequadamente as condições climáticas antecedentes aos eventos de incêndio na Caatinga.

Para isso, foi realizada inicialmente a seleção e associação dos dados meteorológicos provenientes do INMET, seguida da integração com registros históricos de ocorrência de incêndios obtidos por meio do BDQueimadas/INPE. Posteriormente, foram executadas etapas de limpeza, tratamento de valores ausentes, construção de variáveis derivadas e incorporação de dependência temporal por meio de janelas de defasagem.

Adicionalmente, considerando o desbalanceamento natural entre registros com e sem ocorrência de incêndios, foram adotadas estratégias específicas para ampliação da classe minoritária e balanceamento do conjunto de dados, buscando reduzir vieses durante o treinamento dos modelos. Ao final do processo, obteve-se uma base consolidada, estruturada e adequada para aplicação de técnicas de aprendizado supervisionado voltadas à previsão de incêndios florestais.

3.3.1 Seleção e Associação dos Dados Meteorológicos

Os dados meteorológicos utilizados neste trabalho foram obtidos a partir das estações automáticas disponibilizadas pelo INMET, que fornece séries históricas contendo diversas variáveis climáticas com periodicidade diária.

Após a obtenção dos dados, foi realizada uma etapa de seleção dos atributos de interesse, mantendo-se apenas as variáveis meteorológicas utilizadas na construção do modelo preditivo, conforme descrito na Seção 3.2.1. Em seguida, os registros foram organizados e padronizados quanto ao formato das datas, identificação das estações meteorológicas e nomenclatura das variáveis, garantindo uniformidade entre os arquivos provenientes das diferentes estações.

Concluída essa etapa, os dados meteorológicos foram preparados para a associação espacial com os registros de focos de incêndio provenientes do BDQueimadas. Para isso, foi utilizada uma estrutura de busca espacial baseada nos algoritmos cKDTree e BallTree, responsáveis por identificar, de forma eficiente, as

três estações meteorológicas mais próximas de cada foco de calor registrado. A adoção dessas estruturas reduziu significativamente o custo computacional da busca pelos vizinhos mais próximos, tornando o processo de associação espacial mais eficiente.

O procedimento de associação permitiu vincular a cada ocorrência de incêndio as condições meteorológicas observadas nas estações mais próximas, constituindo a base para as etapas de integração dos dados, engenharia de atributos e construção do dataset supervisionado, descritas nas seções seguintes.

3.3.2 Filtragem e Integração dos Dados

Após a seleção das variáveis meteorológicas, foi realizada a filtragem dos registros de incêndios com o objetivo de manter apenas as ocorrências pertencentes ao bioma Caatinga, garantindo alinhamento com o recorte geográfico definido neste estudo. Essa etapa foi necessária para evitar a inclusão de padrões climáticos característicos de outros biomas, que poderiam introduzir ruídos e comprometer a capacidade de generalização do modelo para o contexto ambiental analisado.

Na etapa seguinte, realizou-se a integração entre os dados meteorológicos obtidos junto ao INMET e os registros de focos de incêndio provenientes do BDQueimadas. Esse processo baseou-se na compatibilização temporal entre a data das observações meteorológicas e a data de ocorrência dos focos de incêndio, bem como na associação espacial entre cada ocorrência e as três estações meteorológicas mais próximas de sua localização, conforme procedimento descrito na Seção 3.3.3.

A partir dessa integração foi construída a variável alvo (*target*), utilizada para representar a ocorrência de incêndios no conjunto de dados. Para cada registro meteorológico diário foi atribuído o valor **1** quando identificado pelo menos um foco de incêndio associado à região e à data correspondente, e o valor **0** na ausência de registros.

Além disso, foi construída uma base temporal contínua para cada estação meteorológica, contemplando todos os dias do período analisado, incluindo tanto

datas com ocorrência quanto datas sem ocorrência de incêndios. Essa estratégia foi adotada para preservar a continuidade temporal da base de dados e evitar vieses decorrentes da utilização exclusiva de registros positivos.

A inclusão de exemplos representando ambos os cenários possibilita que os algoritmos de aprendizado supervisionado aprendam padrões discriminativos entre condições ambientais associadas e não associadas à ocorrência de incêndios florestais, contribuindo para maior robustez e capacidade de generalização do modelo preditivo.

3.3.3 Associação Espacial entre Estações Meteorológicas e Focos de Incêndio

Uma das etapas mais importantes da construção do conjunto de dados consistiu na associação espacial entre os registros de focos de incêndio e as informações meteorológicas provenientes das estações automáticas do INMET. Como as estações meteorológicas estão distribuídas de forma irregular no território e os focos de incêndio podem ocorrer em qualquer localização geográfica, foi necessário estabelecer um procedimento que permitisse relacionar cada ocorrência às condições meteorológicas mais representativas de sua região.

Diferentemente de abordagens que utilizam apenas a estação meteorológica mais próxima, neste trabalho adotou-se uma estratégia baseada na associação de cada foco de incêndio às três estações meteorológicas mais próximas. Essa abordagem busca representar de forma mais adequada a variabilidade espacial das condições atmosféricas, reduzindo possíveis vieses decorrentes da utilização de uma única estação de referência.

Para realizar essa associação espacial, inicialmente foram obtidas as coordenadas geográficas (latitude e longitude) de todas as estações meteorológicas e de todos os focos de incêndio. Em seguida, essas coordenadas foram organizadas em estruturas espaciais que permitiram calcular, de forma eficiente, as distâncias entre cada foco de incêndio e as estações meteorológicas disponíveis.

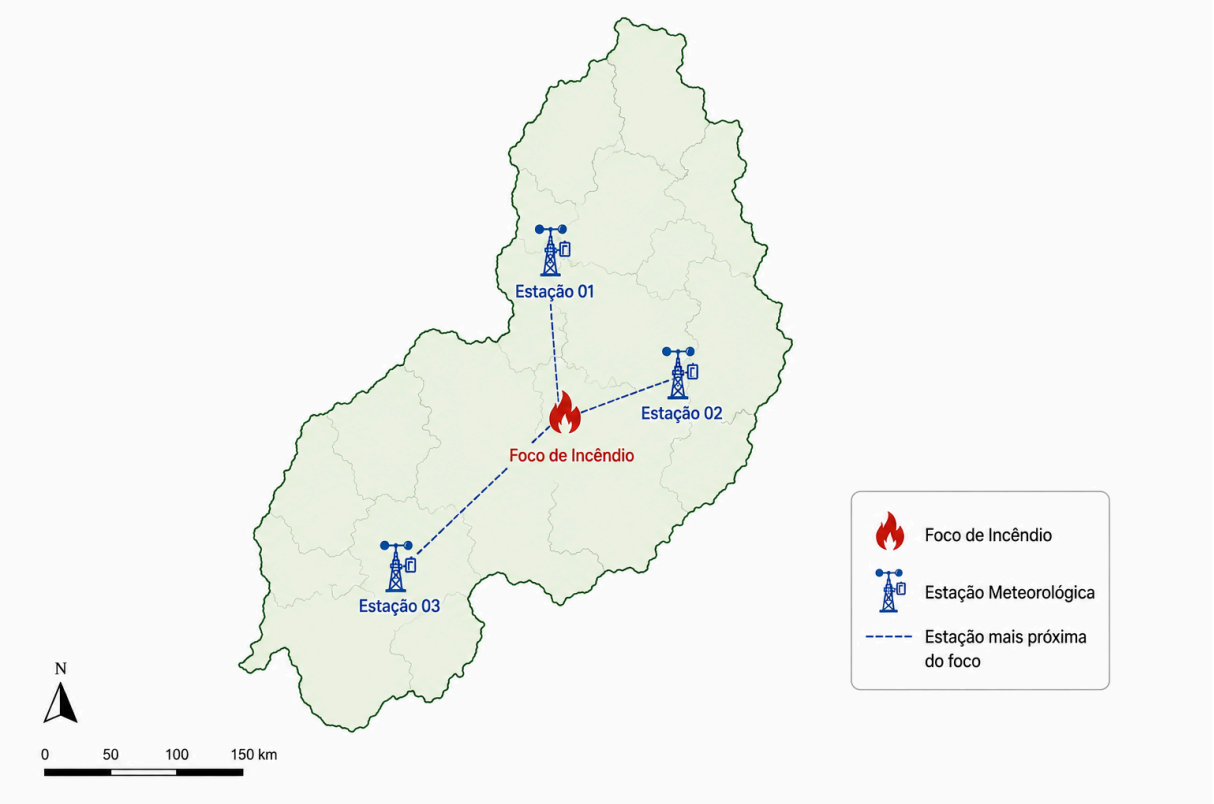
Nesse processo, foram utilizadas as estruturas de busca espacial BallTree e cKDTree, amplamente empregadas em aplicações de geoprocessamento e análise

espacial por possibilitarem consultas rápidas de vizinhança. Essas estruturas organizam os pontos geográficos em árvores de busca, reduzindo significativamente o custo computacional em comparação com a avaliação exaustiva de todas as combinações possíveis entre focos de incêndio e estações meteorológicas.

Após a identificação das três estações mais próximas, foram recuperadas as variáveis meteorológicas correspondentes a cada uma delas e incorporadas ao conjunto de dados, preservando separadamente os atributos da estação mais próxima, da segunda mais próxima e da terceira mais próxima. Dessa forma, cada registro passou a representar não apenas as condições climáticas de uma única estação, mas um conjunto de informações provenientes das estações localizadas nas proximidades do foco de incêndio.

A Figura 5 ilustra o procedimento adotado para a associação espacial entre os focos de incêndio e as estações meteorológicas. Inicialmente é identificado um foco de incêndio registrado pelo BDQueimadas/INPE. Em seguida, são localizadas as três estações meteorológicas mais próximas utilizando a estrutura de busca espacial. Por fim, as variáveis meteorológicas dessas estações são vinculadas ao registro correspondente, compondo uma amostra utilizada na construção do conjunto de dados.

Figura 5 - Associação espacial entre um foco de incêndio e as três estações meteorológicas mais próximas.



Fonte: Elaborado pelo autor (2026)

A adoção desse procedimento constitui um dos principais diferenciais metodológicos desta pesquisa, pois busca aumentar a representatividade espacial das variáveis meteorológicas utilizadas pelo modelo preditivo. Além de reduzir possíveis vieses decorrentes da distribuição geográfica das estações meteorológicas, essa estratégia contribui para representar de maneira mais fiel as condições atmosféricas existentes nas proximidades de cada foco de incêndio, favorecendo o desenvolvimento de modelos com maior capacidade de generalização.

3.3.4 Pareamento Espacial das Observações sem Incêndio

Durante a construção do conjunto de dados foi identificada uma diferença estrutural entre as classes utilizadas no treinamento do modelo. Os registros

correspondentes à ocorrência de incêndios foram obtidos a partir dos focos de calor detectados pelo BDQueimadas, possuindo coordenadas geográficas específicas para cada evento. Em contrapartida, os registros referentes aos dias sem ocorrência de incêndio estavam inicialmente associados apenas às estações meteorológicas, não representando pontos distribuídos ao longo da área de estudo.

Caso essas observações fossem utilizadas diretamente, os registros da classe negativa permaneceriam concentrados nas localizações das estações meteorológicas, enquanto os registros positivos estariam distribuídos espacialmente por diferentes regiões do estado do Piauí. Essa diferença poderia introduzir um viés espacial no processo de treinamento, levando o modelo a aprender padrões relacionados à localização das estações meteorológicas, em vez de identificar as condições climáticas efetivamente associadas à ocorrência de incêndios florestais.

Para reduzir esse problema, foi desenvolvido um procedimento de pareamento espacial destinado à geração das observações sem ocorrência de incêndio. Inicialmente, foi definida uma zona de exclusão com raio de 40 km ao redor de cada foco de calor registrado. Essa etapa teve como objetivo impedir que amostras negativas fossem geradas em regiões muito próximas de incêndios reais, reduzindo a probabilidade de classificar como ausência de incêndio áreas submetidas às mesmas condições ambientais que favoreceram a ocorrência do fogo.

Após a definição da zona de exclusão, foi construída uma malha regular de pontos com espaçamento de 5 km, distribuída por toda a área de estudo. Em seguida, todos os pontos localizados no interior das zonas de exclusão foram removidos, permanecendo apenas aqueles considerados candidatos à geração das amostras negativas. Esse procedimento permitiu distribuir espacialmente as observações da classe negativa, evitando sua concentração exclusivamente nas localizações das estações meteorológicas.

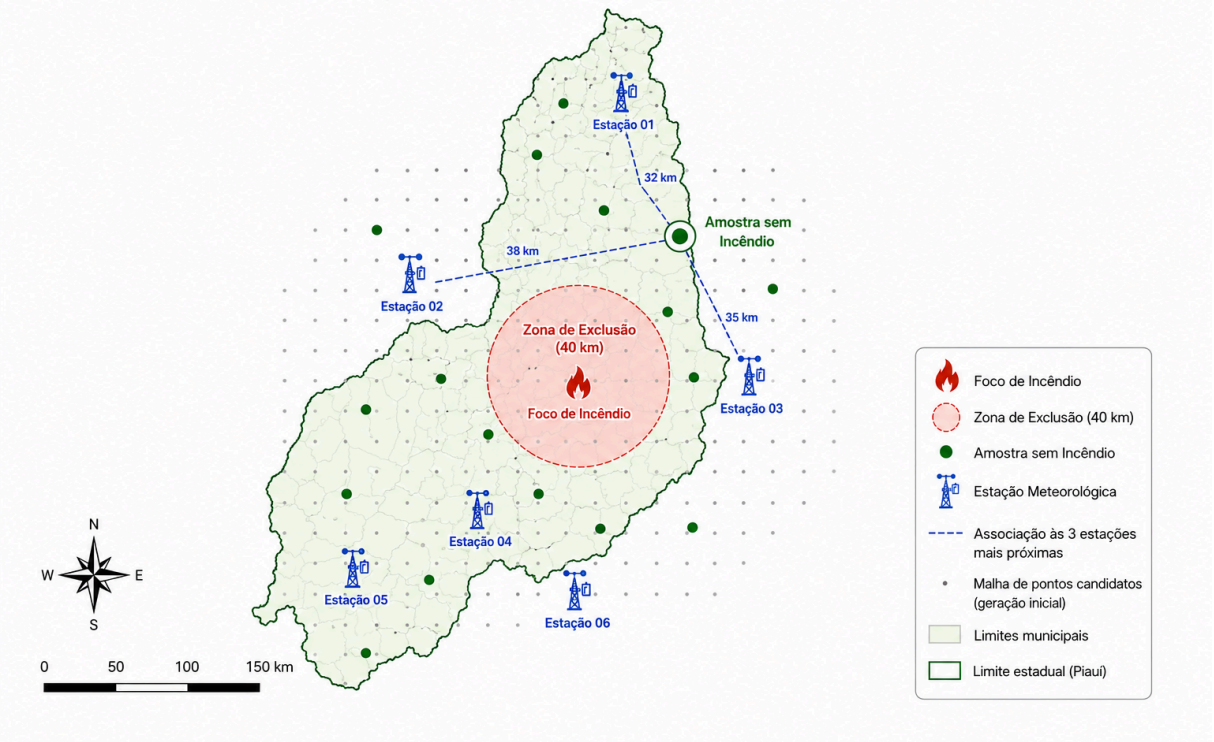
Na sequência, foi utilizada a estrutura de busca espacial cKDTree, responsável por organizar as coordenadas geográficas em uma árvore multidimensional que permite localizar rapidamente os vizinhos mais próximos de

um determinado ponto. Essa estrutura possibilitou identificar, para cada ponto candidato, as três estações meteorológicas mais próximas, reduzindo significativamente o custo computacional quando comparado à avaliação exaustiva das distâncias entre todos os pontos e todas as estações meteorológicas disponíveis.

Após a identificação das três estações mais próximas, foram recuperadas as variáveis meteorológicas correspondentes à estação mais próxima, à segunda mais próxima e à terceira mais próxima, preservando separadamente os atributos de cada uma delas no conjunto de dados. Dessa forma, cada observação sem ocorrência de incêndio passou a representar as condições atmosféricas existentes em seu entorno, seguindo o mesmo procedimento metodológico adotado para os registros com ocorrência de incêndio.

A Figura 6 apresenta o procedimento utilizado para a geração das observações sem ocorrência de incêndio. Inicialmente, é construída uma malha regular de pontos distribuídos sobre a área de estudo. Em seguida, são removidos todos os pontos localizados dentro da zona de exclusão estabelecida ao redor dos focos de incêndio. Para cada ponto remanescente, são identificadas as três estações meteorológicas mais próximas utilizando a estrutura cKDTree. Por fim, as variáveis meteorológicas dessas três estações são incorporadas ao conjunto de dados, originando as amostras negativas utilizadas no treinamento do modelo.

Figura 6 - Processo de geração das amostras sem ocorrência de incêndio.



Fonte: Elaborada pelo autor (2026)

Esse procedimento constitui um dos principais diferenciais metodológicos desta pesquisa, pois busca reduzir vieses decorrentes da distribuição espacial das observações utilizadas no treinamento do modelo. Ao adotar o mesmo critério de associação espacial para ambas as classes, vinculando cada observação, com ou sem ocorrência de incêndio, às três estações meteorológicas mais próximas, obtém-se maior consistência na construção do conjunto de dados.

Essa estratégia contribui para representar de forma mais fiel a variabilidade das condições atmosféricas, reduzindo a influência da distribuição geográfica das estações meteorológicas e favorecendo o aprendizado de padrões climáticos relacionados à ocorrência de incêndios florestais. Como consequência, espera-se aumentar a capacidade de generalização do modelo e ampliar sua aplicação em diferentes regiões do estado do Piauí.

3.3.5 Organização Temporal e Criação de Variáveis Derivadas

Após a integração entre os dados meteorológicos e os registros de incêndios, o conjunto de dados foi organizado cronologicamente e ordenado por estação meteorológica e data de observação. Essa etapa preserva a sequência temporal das informações, considerando que a ocorrência de incêndios florestais é influenciada pelo acúmulo de condições ambientais favoráveis ao longo do tempo.

Com base nessa organização, foram geradas variáveis derivadas destinadas a representar características climáticas que não são capturadas diretamente pelos registros diários. Entre elas, destaca-se a variável dias sem chuva, que contabiliza o número de dias consecutivos sem precipitação e atua como um indicador indireto da intensidade da estiagem e do ressecamento progressivo da vegetação.

Também foram incorporadas variáveis temporais, como mês e dia do ano, permitindo representar a sazonalidade do regime climático e identificar períodos historicamente mais suscetíveis à ocorrência de incêndios florestais.

Além disso, foram preservadas as variáveis geográficas associadas às estações meteorológicas, incluindo latitude, longitude e altitude. Esses atributos permitem representar diferenças espaciais e topográficas que influenciam as condições climáticas locais e, consequentemente, a suscetibilidade à ocorrência de incêndios.

Ao final dessa etapa, o conjunto de dados passou a ser composto por atributos meteorológicos, temporais, geográficos e derivados, fornecendo uma representação mais abrangente das condições ambientais relacionadas ao fenômeno estudado.

3.3.6 Construção da Janela Temporal

Considerando que a ocorrência de incêndios florestais é influenciada não apenas pelas condições meteorológicas do dia analisado, mas também pelo comportamento climático dos dias anteriores, foi adotada uma abordagem baseada em janelas temporais. Essa estratégia permite incorporar ao conjunto de dados informações sobre a persistência das condições ambientais que favorecem a ignição e a propagação do fogo.

Para isso, foram criadas variáveis de defasagem temporal (lags) referentes aos três dias anteriores à data de cada observação, representadas por D−1, D−2 e D−3. Esse procedimento foi aplicado às variáveis de precipitação diária, pressão atmosférica média, temperatura média do ar, umidade relativa média e velocidade média do vento.

A utilização dessas variáveis possibilita representar os efeitos acumulados das condições meteorológicas recentes, como períodos consecutivos de estiagem, manutenção de temperaturas elevadas, baixa umidade relativa do ar e persistência de ventos favoráveis à propagação dos incêndios.

A definição de uma janela temporal de três dias buscou representar os efeitos climáticos de curto prazo sem aumentar excessivamente a dimensionalidade do conjunto de dados, reduzindo a inclusão de atributos redundantes e preservando a eficiência computacional.

Após a geração das variáveis de defasagem, os registros que apresentavam valores ausentes decorrentes da inexistência de observações dos dias anteriores foram removidos, assegurando a consistência temporal do conjunto de dados utilizado nas etapas subsequentes do estudo.

Ao término das etapas de integração, engenharia de atributos e construção da janela temporal, o conjunto de dados passou a ser composto por variáveis meteorológicas, espaciais, temporais e derivadas. O **Quadro 2** apresenta a organização das variáveis utilizadas na construção do conjunto de dados, agrupadas conforme sua finalidade e origem.

Quadro 2: Lista de variáveis do dataset

Categoria	**Variáveis**
Variáveis meteorológicas originais	Precipitação diária, pressão atmosférica média, temperatura do ponto de orvalho, temperatura máxima,

	temperatura média, temperatura mínima, umidade relativa média, umidade relativa mínima, rajada máxima do vento e velocidade média do vento.
Variáveis temporais	Data, mês e dia do ano.
Variáveis espaciais	Estação meteorológica associada, distância até a estação principal, distância até a segunda estação e distância até a terceira estação.
Variáveis derivadas	Dias sem chuva, precipitação acumulada nos últimos 7 dias, temperatura média dos últimos 7 dias e umidade média dos últimos 7 dias.
Variáveis de defasagem (*lag*)	Valores de precipitação, temperatura média, umidade relativa média, velocidade do vento e pressão atmosférica referentes aos 1, 2 e 3 dias anteriores.
Variáveis das estações vizinhas	Informações meteorológicas das três estações meteorológicas vizinhas mais próximas (precipitação, temperatura média, umidade relativa, velocidade do vento, pressão atmosférica e dias sem chuva).
Variável alvo	Ocorrência de incêndio (*incêndio*), utilizada como variável resposta na classificação binária.

Fonte: Elaborado pelo autor (2026)

3.3.7 Tratamento de Valores Ausentes e Limpeza dos Dados

Durante o pré-processamento dos dados, foram identificadas lacunas em algumas variáveis meteorológicas, decorrentes principalmente de falhas de registro nas estações automáticas e da indisponibilidade de medições em determinados períodos.

Como etapa inicial de controle de qualidade, foi realizada uma análise da completude dos registros. Observações com mais de 50% de dados ausentes foram removidas da base, uma vez que elevados níveis de incompletude poderiam comprometer a representatividade das amostras e afetar a confiabilidade das análises subsequentes.

Para os registros que apresentavam percentual de ausência inferior a esse limite, adotou-se uma estratégia de imputação baseada na interpolação temporal simples. Os valores faltantes foram estimados por meio da média aritmética entre as observações registradas no dia imediatamente anterior e no dia imediatamente posterior, preservando a continuidade das séries temporais e reduzindo a perda de informações.

Esse procedimento é amplamente empregado em séries temporais meteorológicas, pois variáveis contínuas, como temperatura, pressão atmosférica e umidade relativa do ar, tendem a apresentar variações graduais em intervalos diários. Assim, a interpolação temporal constitui uma alternativa adequada para minimizar o impacto de pequenas lacunas sem alterar significativamente o comportamento das séries.

Após a imputação dos dados, foi realizada uma verificação final de consistência para assegurar que o conjunto de dados estivesse livre de registros inconsistentes ou incompletos, tornando-o adequado para as etapas de modelagem e avaliação do modelo preditivo.

3.3.8 Balanceamento do Dataset

Foi observado que o *dataset* apresenta desbalanceamento entre as classes, com predominância de registros correspondentes a dias sem ocorrência de incêndio. Esse comportamento é esperado em problemas de previsão de eventos raros, uma

vez que os incêndios representam apenas uma pequena parcela do total de observações disponíveis.

Como estratégia para aumentar a representatividade da classe minoritária, realizou-se a ampliação da base histórica por meio da incorporação de registros de incêndios referentes ao período de 2010 a 2015, aumentando o número de exemplos positivos disponíveis para treinamento. Ainda assim, o conjunto de dados permaneceu naturalmente desbalanceado, característica frequentemente observada em problemas de previsão de eventos extremos.

Dessa forma, durante o treinamento do modelo foi utilizado o parâmetro **class_weight = "balanced"** do algoritmo *LightGBM*. Esse recurso ajusta automaticamente os pesos atribuídos às classes de acordo com sua frequência no conjunto de treinamento, atribuindo maior penalização aos erros cometidos sobre a classe minoritária.

A utilização dessa estratégia reduz a tendência do modelo em favorecer a classe majoritária, aumentando sua capacidade de identificar corretamente ocorrências de incêndios florestais sem a necessidade de técnicas de sobreamostragem (*oversampling*) ou geração sintética de novas amostras.

3.3.9 Divisão do Conjunto de Dados

Após as etapas de integração, limpeza, tratamento dos dados e a produção de novos atributos, o *dataset* final foi dividido em subconjuntos destinados ao treinamento, validação e teste do modelo preditivo.

A divisão foi realizada por meio de amostragem estratificada (*stratified sampling*), preservando a proporção original entre as classes em todos os subconjuntos. Além disso, foi utilizada uma semente aleatória fixa, garantindo a reprodutibilidade dos experimentos.

Inicialmente, 70% dos registros foram destinados ao conjunto de treinamento, enquanto os 30% restantes foram reservados para validação e teste. Em seguida,

essa parcela foi subdividida igualmente, resultando em aproximadamente 15% dos dados para validação e 15% para teste final.

O conjunto de treinamento foi utilizado para o ajuste dos parâmetros do algoritmo *LightGBM*. Durante esse processo, o conjunto de validação foi empregado para monitorar o desempenho do modelo e aplicar o mecanismo de *early stopping*, interrompendo o treinamento quando não eram observadas melhorias após um número pré-definido de iterações, reduzindo assim o risco de sobreajuste (*overfitting*).

Além disso, as probabilidades produzidas pelo modelo no conjunto de validação foram utilizadas para determinar o limiar de classificação (*threshold*). Para isso, foi construída a curva *Precision–Recall* e selecionado o valor que maximizou o F1-Score, buscando um equilíbrio entre precisão (*precision*) e revocação (*recall*) na classificação das ocorrências de incêndios florestais.

Por fim, o conjunto de teste permaneceu completamente isolado durante todas as etapas de treinamento e ajuste do modelo, sendo utilizado exclusivamente para a avaliação final de sua capacidade de generalização. Essa avaliação foi realizada por meio de métricas de desempenho amplamente empregadas em problemas de classificação, incluindo AUC-ROC, Precisão (*Precision*), Revocação (*Recall*), F1-Score, Coeficiente Kappa de Cohen (*Cohen's Kappa*) e Matriz de Confusão.

3.4 CONSTRUÇÃO E MÉTRICAS DE AVALIAÇÃO DO MODELO

Após a preparação e organização do *dataset*, foi desenvolvido o modelo preditivo utilizando o algoritmo *LightGBM*, uma implementação otimizada da técnica de *Gradient Boosting* baseada em árvores de decisão. A escolha desse algoritmo fundamenta-se em sua elevada eficiência computacional, boa capacidade de generalização e desempenho em problemas de classificação envolvendo dados tabulares, além de sua habilidade em modelar relações não lineares entre as variáveis de entrada e a variável de saída.

O *LightGBM* constrói o modelo de forma iterativa, adicionando

sucessivamente novas árvores de decisão que buscam corrigir os erros cometidos pelas árvores anteriormente geradas. Diferentemente de algoritmos tradicionais baseados em árvores, o *LightGBM* emprega uma estratégia de crescimento por folhas (*leaf-wise growth*), permitindo melhor aproveitamento da estrutura dos dados e, frequentemente, maior desempenho preditivo.

O modelo foi implementado utilizando a biblioteca *LightGBM* para Python, sendo configurado com 1000 árvores de decisão (*n_estimators = 1000*), taxa de aprendizado igual a 0,05 (*learning_rate = 0.05*) e semente aleatória fixa (*random_state = 42*), garantindo a reprodutibilidade dos experimentos.

Considerando o desbalanceamento existente entre os registros com e sem ocorrência de incêndios, foi utilizado o parâmetro class_weight = "balanced", responsável por ajustar automaticamente os pesos das classes durante o treinamento. Essa estratégia reduz o viés em favor da classe majoritária e aumenta a capacidade do modelo em identificar corretamente os eventos de incêndio.

Durante o treinamento foi empregada a técnica de Early Stopping, utilizando o conjunto de validação para monitorar continuamente o desempenho do modelo. O treinamento era interrompido automaticamente quando não eram observadas melhorias durante cinquenta iterações consecutivas (*stopping_rounds = 50*), reduzindo o risco de sobreajuste (*overfitting*) e contribuindo para a obtenção de um modelo com maior capacidade de generalização.

Após o treinamento, o modelo passou a estimar, para cada observação de entrada, a probabilidade de ocorrência de incêndio. Essas probabilidades foram utilizadas posteriormente na definição do limiar de classificação mais adequado para o problema investigado.

3.4.1 Otimização de Limiar de Decisão

Como o modelo *LightGBM* fornece probabilidades de pertencimento à classe positiva, foi realizada uma etapa de definição do limiar (*threshold*) de classificação, buscando otimizar o desempenho do modelo em um cenário caracterizado pelo desbalanceamento entre as classes.

Inicialmente, o modelo foi aplicado ao conjunto de validação, obtendo-se as probabilidades estimadas para cada observação. Em seguida, foi construída a curva Precision–Recall, da qual foram obtidos automaticamente os diferentes valores candidatos de *threshold*. Para cada limiar, calculou-se o F1-Score, métrica que combina precisão (*Precision*) e sensibilidade (*Recall*), sendo especialmente indicada para problemas de classificação com classes desbalanceadas.

O limiar correspondente ao maior valor de F1-Score foi selecionado como *threshold* ótimo e posteriormente aplicado na classificação dos exemplos do conjunto de teste; a partir daí foram aferidas as métricas finais de avaliação do modelo.

A utilização dessa estratégia permitiu adequar o processo de classificação ao objetivo do trabalho, priorizando um equilíbrio entre a identificação correta de ocorrências de incêndio e a redução de classificações incorretas, tornando o modelo mais apropriado para aplicações voltadas ao monitoramento e ao apoio à tomada de decisão.

3.4.2 Métricas de Avaliação do Desempenho do Modelo

Após o treinamento e a definição do limiar de classificação, o desempenho do modelo foi avaliado utilizando o conjunto de teste, composto por dados que permaneceram completamente isolados durante as etapas anteriores. Essa estratégia permite obter uma estimativa mais confiável da capacidade de generalização do modelo quando aplicado a dados não utilizados em seu processo de aprendizagem.

A avaliação foi realizada por meio de diferentes métricas amplamente empregadas em problemas de classificação supervisionada. Inicialmente, foi gerado o Relatório de Classificação (*Classification Report*), contendo as métricas de Precisão (Precision), Sensibilidade (Recall), F1-Score e Acurácia para cada classe analisada. A métrica *Precision* indica a proporção de previsões positivas corretamente classificadas, enquanto o *Recall* representa a capacidade do modelo em identificar corretamente os registros correspondentes à ocorrência de incêndios. O F1-Score, por sua vez, corresponde à média harmônica entre precisão e

sensibilidade, sendo particularmente adequado para bases de dados desbalanceadas, como a utilizada neste trabalho.

Além dessas métricas, foi calculada a Área sob a Curva ROC (AUC-ROC), utilizada para mensurar a capacidade discriminativa do modelo em diferentes limiares de classificação. Valores mais elevados dessa métrica indicam maior capacidade do classificador em distinguir corretamente registros pertencentes às classes "incêndio" e "não incêndio".

Também foi utilizado o Coeficiente Kappa de Cohen, métrica que avalia o grau de concordância entre as classificações realizadas pelo modelo e os valores reais, descontando a concordância esperada ao acaso. Essa medida fornece uma avaliação complementar do desempenho do modelo, especialmente em conjuntos de dados com distribuição desigual entre as classes.

Por fim, foi construída a Matriz de Confusão, permitindo visualizar a distribuição das classificações corretas e incorretas realizadas pelo modelo, identificando a quantidade de verdadeiros positivos, verdadeiros negativos, falsos positivos e falsos negativos. Adicionalmente, foi gerada a Curva Precision–Recall, utilizada para analisar o comportamento do modelo em diferentes níveis de sensibilidade e precisão, oferecendo uma avaliação mais apropriada para problemas de classificação com classes desbalanceadas.

3.4.3 Aplicação do Modelo para Geração de Mapas de Risco

Após a etapa de treinamento, avaliação e armazenamento do modelo preditivo, foi desenvolvida uma aplicação destinada à utilização prática do sistema em cenários reais de monitoramento ambiental. Essa aplicação tem como objetivo estimar o risco de ocorrência de incêndios florestais para uma data específica e representar espacialmente os resultados por meio de mapas de calor.

Inicialmente, o modelo previamente treinado foi carregado utilizando a biblioteca **Joblib**, juntamente com o limiar ótimo de classificação (*threshold*) e a relação das variáveis empregadas durante o treinamento. Em seguida, foram fornecidos ao modelo os dados meteorológicos correspondentes ao período de

interesse, contendo as mesmas variáveis utilizadas durante a etapa de aprendizagem.

Para cada registro analisado, o algoritmo *LightGBM* estimou a probabilidade de ocorrência de incêndio por meio da função predict_proba(). As probabilidades obtidas representam o grau de suscetibilidade à ocorrência de incêndios considerando as condições meteorológicas observadas para cada estação.

Posteriormente, essas estimativas probabilísticas foram associadas às respectivas coordenadas geográficas das estações meteorológicas e submetidas a um processo de interpolação espacial, permitindo estimar continuamente o comportamento do risco entre os pontos observados. O resultado desse procedimento consiste em um mapa de calor, no qual diferentes intensidades de cor representam diferentes níveis de risco previstos pelo modelo.

Essa representação espacial possibilita identificar regiões potencialmente mais suscetíveis à ocorrência de incêndios florestais, oferecendo uma ferramenta de apoio à tomada de decisão para ações de monitoramento, prevenção e planejamento operacional. Além disso, a geração dos mapas demonstra a aplicabilidade prática da metodologia proposta, permitindo transformar as previsões numéricas produzidas pelo modelo em informações espaciais de fácil interpretação pelos órgãos responsáveis pela gestão ambiental.

3.5 ASPECTOS ÉTICOS DA PESQUISA

A presente pesquisa caracteriza-se como um estudo computacional baseado na análise de dados públicos disponibilizados por órgãos oficiais, não envolvendo experimentação com seres humanos, animais ou organismos vivos. Dessa forma, não há riscos físicos, psicológicos, sociais ou morais decorrentes de sua execução, não sendo necessária submissão ao Comitê de Ética em Pesquisa.

Todos os dados utilizados foram obtidos em bases públicas disponibilizadas pelo INMET e pelo BDQueimadas, respeitando os princípios de transparência, legalidade e acesso público à informação.

Durante todas as etapas do trabalho foram adotadas boas práticas científicas,

incluindo a descrição detalhada da metodologia empregada, a utilização adequada das fontes de dados, a reprodutibilidade dos experimentos computacionais e a apresentação fidedigna dos resultados obtidos, assegurando a integridade científica da pesquisa.

4 RESULTADOS E DISCUSSÃO

4.1 ANÁLISE E OTIMIZAÇÃO DO LIMIAR DE DECISÃO (*THRESHOLD*)

O modelo *LightGBM* gera, para cada observação, uma probabilidade de ocorrência de incêndio florestal, representando a estimativa de pertencimento à classe positiva. Em problemas de classificação binária, é comum utilizar o limiar de decisão (threshold) igual a 0,5 para converter essas probabilidades em classes. Entretanto, essa abordagem pode não ser a mais adequada em aplicações de monitoramento ambiental, nas quais a não detecção de um evento real pode gerar consequências mais significativas do que a emissão de um alerta falso.

Com o objetivo de obter um limiar de classificação mais apropriado ao problema estudado, foi utilizada a base de validação para analisar diferentes valores de threshold. Para cada limiar avaliado foram calculadas as métricas de precisão, recall e F1-Score, sendo selecionado aquele que apresentou o maior valor de F1-Score para a classe positiva.

O processo resultou na definição de um limiar ótimo de 0,43, posteriormente aplicado ao conjunto de teste. A Tabela 1 apresenta a comparação entre o desempenho do modelo utilizando o limiar padrão (threshold = 0,50) e o limiar otimizado (threshold = 0,43).

Observa-se que a redução do limiar aumentou o recall da classe positiva de 0,84 para 0,88, indicando maior capacidade do modelo em identificar corretamente as ocorrências reais de incêndio florestal. Em contrapartida, a precisão da mesma classe apresentou uma pequena redução, passando de 0,81 para 0,79, o que representa um aumento moderado na quantidade de falsos positivos. Apesar dessa variação, o F1-Score da classe positiva permaneceu em 0,83, evidenciando que o equilíbrio entre precisão e sensibilidade foi mantido.

As métricas globais do modelo permaneceram praticamente inalteradas após o ajuste do limiar. A acurácia manteve-se em 0,82, a AUC-ROC permaneceu em 0,8968, e o coeficiente de Cohen's Kappa apresentou apenas uma pequena variação, de 0,6425 para 0,6382, indicando que o desempenho geral do classificador foi preservado.

Assim, a utilização de um limiar ajustado mostrou-se mais adequada ao contexto desta pesquisa do que a adoção do valor padrão de 0,5. O aumento da sensibilidade do modelo possibilitou identificar uma quantidade maior de ocorrências reais de incêndio, característica desejável em aplicações de monitoramento ambiental, nas quais a redução de falsos negativos é prioritária para apoiar ações preventivas e a tomada de decisão pelos órgãos responsáveis.

Tabela 1: Comparação do Threshold

Métrica	Threshold = 0,50	Threshold = 0,43
Precisão (classe 1)	0,81	0,79
Recall (classe 1)	0,84	0,88
F1-Score (classe 1)	0,83	0,83
Acurácia	0,82	0,82
AUC-ROC	0,89	0,89
Cohen's Kappa	0,64	0,63

Fonte: Elaborado pelo autor (2026)

4.2 AVALIAÇÃO DE DESEMPENHO DA CLASSIFICAÇÃO

Após a definição do limiar ótimo de classificação, o modelo foi avaliado utilizando o conjunto de teste, composto por registros que não participaram das etapas de treinamento e validação. Os resultados obtidos são apresentados na Tabela 2.

Tabela 2: Métricas adquiridas pelo modelo

Métricas	Valor
Acurácia	82,0%
Precisão	79,0%
Recall	88,0%
F1 - Score	83,0%
AUC-ROC	89,7%
Coeficiente Kappa	63,8%

Fonte: Elaborado pelo autor (2026)

Os resultados demonstram que o modelo apresentou bom desempenho na classificação das ocorrências de incêndio. A acurácia de 82% indica elevada proporção de classificações corretas. Entretanto, considerando que o objetivo da pesquisa é a identificação preventiva de incêndios florestais, as métricas de precisão, recall e F1-Score fornecem uma avaliação mais adequada do desempenho do modelo.

A precisão de 79% indica que aproximadamente quatro em cada cinco alertas emitidos pelo modelo corresponderam a ocorrências reais de incêndio. Por sua vez, o recall de 88% evidencia elevada capacidade de identificar corretamente eventos de incêndio, reduzindo a quantidade de falsos negativos, aspecto de grande importância para aplicações voltadas ao monitoramento preventivo.

O F1-Score de 83% demonstra um bom equilíbrio entre precisão e sensibilidade, enquanto a AUC-ROC de 90% indica excelente capacidade discriminativa, evidenciando que o modelo consegue distinguir adequadamente observações com e sem ocorrência de incêndio. Além disso, o Coeficiente Kappa de Cohen (0,6382) representa uma concordância substancial entre as classificações produzidas pelo modelo e os registros observados.

Como forma de contextualizar os resultados obtidos, a Tabela 3 apresenta uma comparação entre o modelo desenvolvido nesta pesquisa e trabalhos recentes encontrados na literatura.

Tabela 3 — Confronto comparativo de desempenho preditivo entre o modelo proposto e a literatura.

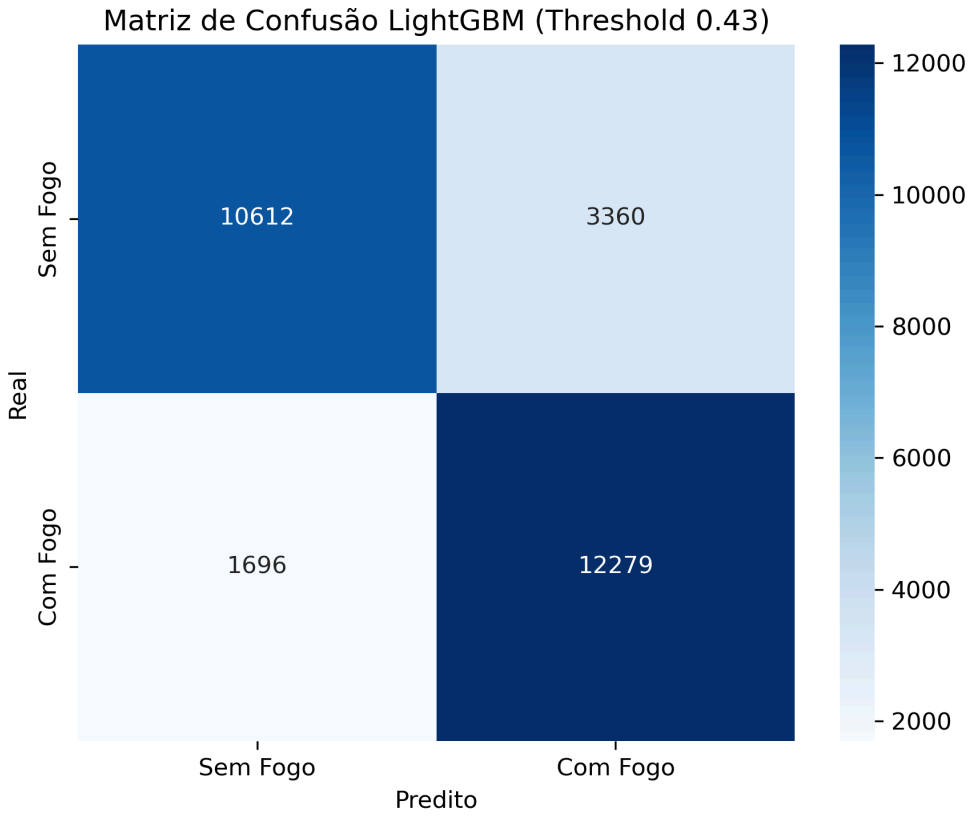
Métrica Operacional	**Modelo Proposto**	**Hanamoto (2025) (Random Forest Tradicional)**
Acurácia Global	82,0%	81,4%
Área Sob a Curva (AUC)	89,7%	88,7%
Precisão (Classe 1)	79,0%	79,4%
Sensibilidade (Recall)	88,0%	83,9%
F1-Score (Classe 1)	83,0%	81,6%

Fonte: Elaborado pelo autor (2026).

A comparação evidencia que o modelo desenvolvido apresentou desempenho competitivo em relação aos trabalhos de referência. Embora a precisão tenha sido semelhante à obtida por Hanamoto (2025), o modelo proposto alcançou valores superiores de *recall* e F1-Score, indicando maior capacidade de identificar corretamente ocorrências de incêndio. Esse comportamento é especialmente relevante em aplicações preventivas, nas quais a redução de falsos negativos possui maior importância do que pequenas perdas de precisão.

A Figura 7 apresenta a matriz de confusão obtida para o conjunto de teste.

Figura 7 - Matriz de confusão

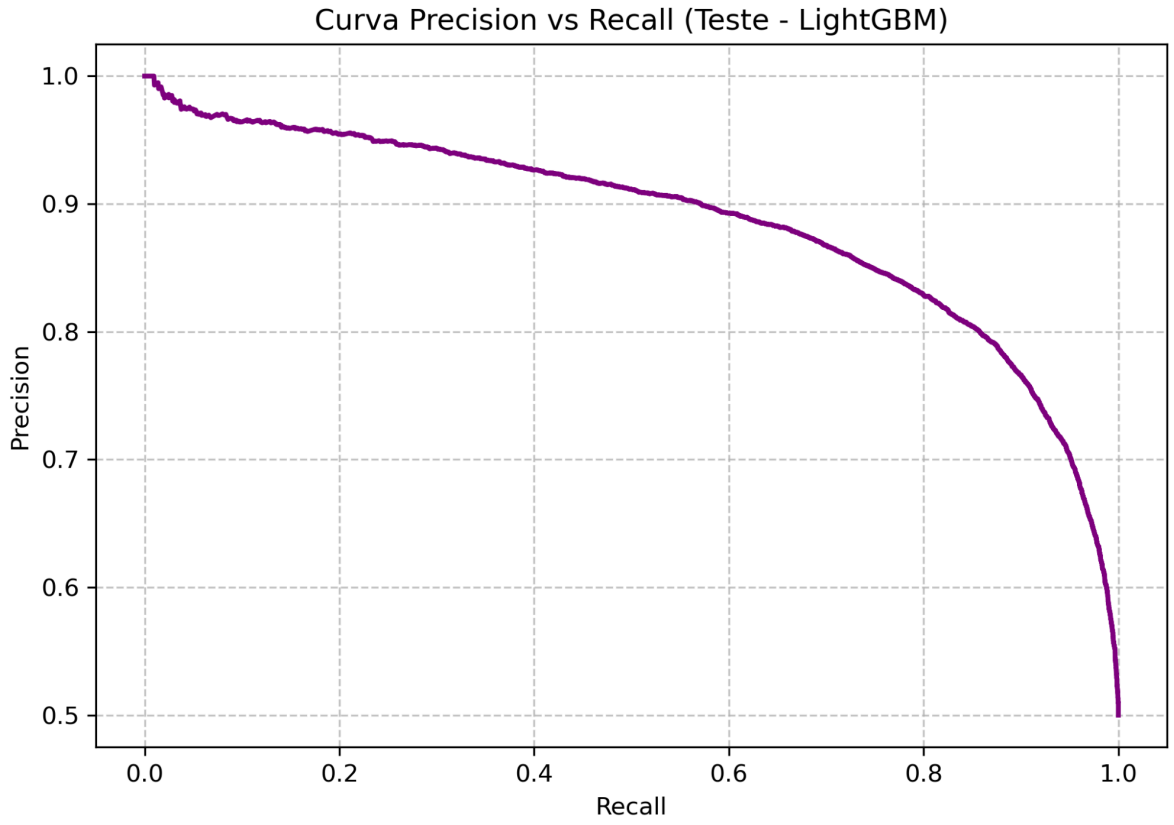


Fonte: Gerado pelo autor (2026) a partir do modelo

A matriz de confusão permite visualizar a distribuição entre verdadeiros positivos, verdadeiros negativos, falsos positivos e falsos negativos. Observa-se predominância de classificações corretas em ambas as classes, corroborando os valores obtidos pelas métricas de desempenho. Destaca-se ainda a reduzida quantidade de falsos negativos, evidenciando a capacidade do modelo em identificar a maior parte dos eventos reais de incêndio.

Complementando essa análise, a Figura 8 apresenta a curva Precision–Recall obtida para o conjunto de teste.

Figura 8 - Precision vs Recall



Fonte: Gerado pelo autor (2026) a partir do modelo.

A curva Precision–Recall demonstra o comportamento conjunto entre precisão e sensibilidade para diferentes valores de limiar de decisão. Esse tipo de análise é particularmente indicado para problemas de classificação em que a correta identificação da classe positiva é prioritária. Os resultados obtidos evidenciam um equilíbrio satisfatório entre as duas métricas, corroborando a escolha do limiar de classificação adotado nesta pesquisa e reforçando a capacidade do modelo em detectar ocorrências de incêndio com boa confiabilidade.

4.3 AVALIAÇÃO DE DESEMPENHO ESTIMATIVO (MAPA DE RISCO DE INCÊNDIO)

Após a etapa de validação, o modelo treinado foi salvo utilizando a biblioteca Joblib, juntamente com o limiar de classificação e a lista de variáveis empregadas durante o treinamento.

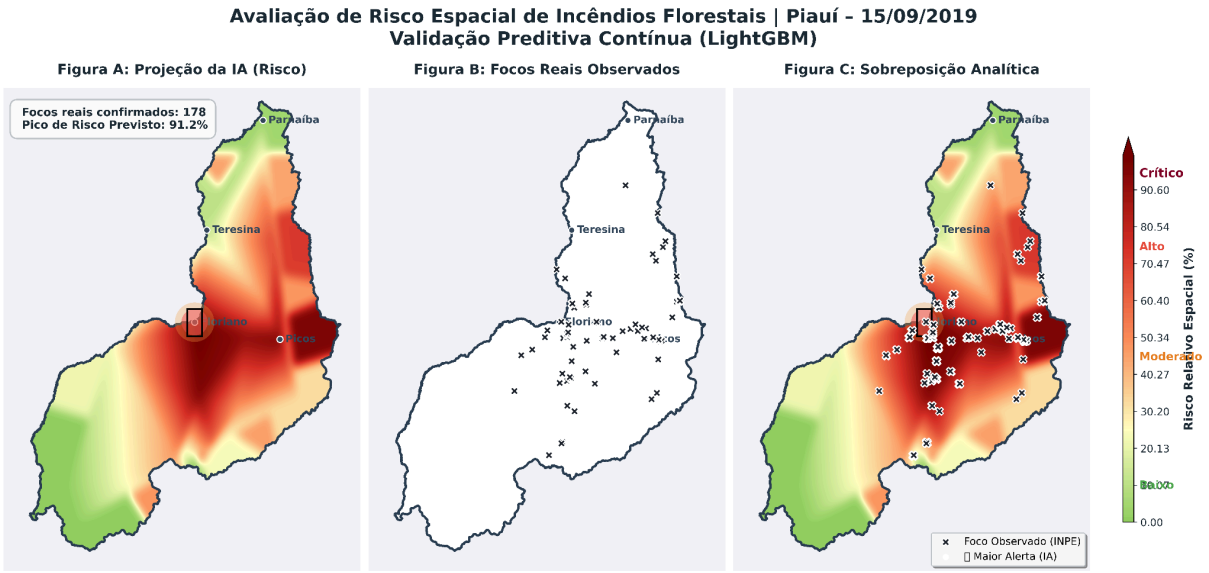
Posteriormente, esses arquivos foram utilizados em uma aplicação desenvolvida para realizar inferências sobre novos dados meteorológicos. A partir das probabilidades estimadas pelo modelo, foi possível gerar mapas espaciais representando o risco relativo de ocorrência de incêndios florestais para uma determinada data.

A Figura 9 apresenta um exemplo da aplicação do modelo para o Estado do Piauí na data de 15 de setembro de 2019. O mapa foi construído a partir das probabilidades de ocorrência de incêndio estimadas pelo modelo LightGBM para cada ponto da grade espacial utilizada no estudo. Essas probabilidades foram posteriormente interpoladas, produzindo uma superfície contínua que representa a distribuição espacial do risco em todo o estado.

A escala de cores indica diferentes níveis de suscetibilidade à ocorrência de incêndios florestais. Tonalidades mais frias representam áreas de menor risco, enquanto tonalidades mais quentes correspondem a regiões com maior probabilidade estimada de ocorrência de incêndios.

Além disso, os focos de calor registrados pelo BDQueimadas/INPE na mesma data foram sobrepostos ao mapa, permitindo uma comparação visual entre as áreas classificadas pelo modelo como mais suscetíveis e os eventos efetivamente observados. Essa comparação possibilita avaliar, de forma qualitativa, a coerência entre a distribuição espacial do risco estimado e os registros históricos de focos de calor.

Figura 9 - Mapa de risco relativo espacial de focos de calor gerado pelo



modelo *LightGBM* para o Estado do Piauí (15/09/2019).

Fonte: Elaborado pelo autor (2026)

A Figura 9 representa o principal produto da metodologia desenvolvida, transformando as probabilidades estimadas pelo modelo em uma representação espacial de fácil interpretação. Diferentemente das métricas de desempenho apresentadas nas seções anteriores, que avaliam a qualidade da classificação por meio de indicadores numéricos, o mapa de calor permite visualizar a distribuição geográfica do risco de ocorrência de incêndios florestais, facilitando a identificação de áreas mais suscetíveis.

Observa-se a formação de regiões contínuas de maior risco, indicando que a suscetibilidade estimada pelo modelo apresenta um comportamento espacial coerente com as condições meteorológicas predominantes na data analisada. Essa distribuição evidencia que o modelo foi capaz de capturar padrões espaciais associados à ocorrência dos incêndios, em vez de produzir previsões isoladas ou aleatórias.

A sobreposição dos focos de calor registrados pelo BDQueimadas/INPE demonstra, de forma qualitativa, que uma parcela significativa dos eventos observados ocorreu em áreas classificadas pelo modelo como de maior risco.

Embora essa análise não substitua as métricas quantitativas de avaliação, ela fornece uma evidência visual da consistência das previsões realizadas, reforçando a capacidade do modelo em representar espacialmente a suscetibilidade à ocorrência de incêndios florestais.

Além da avaliação do desempenho, o mapa evidencia o potencial da metodologia como ferramenta de apoio ao monitoramento ambiental. A representação espacial do risco pode auxiliar os órgãos responsáveis pela prevenção e combate aos incêndios na identificação de áreas prioritárias para monitoramento, no planejamento de ações preventivas e na otimização da alocação de recursos humanos e operacionais. Dessa forma, o modelo desenvolvido não apenas apresenta bom desempenho na classificação dos eventos, mas também produz um resultado de aplicação prática, capaz de subsidiar a tomada de decisão em atividades de gestão ambiental.

5 CONCLUSÕES E TRABALHOS FUTUROS

O presente trabalho teve como objetivo desenvolver e avaliar um modelo preditivo baseado em Aprendizado de Máquina para estimar a ocorrência de incêndios florestais no estado do Piauí a partir de variáveis meteorológicas. Para isso, foram integrados registros históricos de focos de calor obtidos por sensoriamento remoto com dados provenientes de estações meteorológicas automáticas do INMET, possibilitando a construção de um conjunto de dados adequado ao treinamento do modelo.

Os resultados obtidos demonstraram que a utilização do algoritmo *LightGBM* apresentou desempenho satisfatório para o problema estudado. O modelo alcançou acurácia de 82,00%, precisão de 79,00%, recall de 88,00%, F1-Score de 83,00%, AUC-ROC de 89,68 e coeficiente Kappa de 63,82%, indicando boa capacidade de distinguir dias com e sem ocorrência de incêndios florestais. Entre as métricas avaliadas, destaca-se o elevado valor de recall, que evidencia a capacidade do modelo em identificar corretamente a maior parte dos eventos de incêndio, característica importante para aplicações voltadas ao monitoramento e à prevenção.

A metodologia proposta também demonstrou que a integração entre dados meteorológicos e registros históricos de focos de calor constitui uma alternativa viável para a construção de sistemas preditivos de baixo custo operacional, utilizando exclusivamente bases de dados públicas. Além disso, a geração de mapas de risco a partir das probabilidades produzidas pelo modelo evidencia o potencial de aplicação prática da abordagem, fornecendo informações que podem auxiliar órgãos ambientais e equipes responsáveis pelo monitoramento e combate aos incêndios florestais.

Entretanto, algumas limitações devem ser consideradas. O modelo foi desenvolvido utilizando apenas variáveis meteorológicas e informações derivadas dessas variáveis, não contemplando fatores como cobertura vegetal, uso e ocupação do solo, relevo, características da vegetação e variáveis socioeconômicas, que também influenciam a ocorrência e a propagação dos incêndios florestais. Além disso, o modelo foi calibrado utilizando dados referentes ao estado do Piauí, sendo necessária uma nova etapa de treinamento, avaliação e análises adicionais para sua aplicação em outras regiões com características climáticas distintas.

Como trabalhos futuros, recomenda-se ampliar a base temporal utilizada no treinamento do modelo, incorporar novas variáveis ambientais, realizar comparações com outros algoritmos de Aprendizado de Máquina e desenvolver uma aplicação web capaz de gerar mapas de risco automaticamente a partir de dados meteorológicos atualizados. Também se mostra interessante integrar o sistema a plataformas de monitoramento ambiental, permitindo sua utilização como ferramenta de apoio à tomada de decisão por instituições responsáveis pela prevenção e combate aos incêndios florestais.

Conclui-se que o uso de técnicas de Aprendizado de Máquina, em especial do algoritmo *LightGBM*, apresenta potencial para auxiliar na previsão de incêndios florestais no estado do Piauí. Os resultados obtidos demonstram que a metodologia proposta é capaz de fornecer estimativas consistentes do risco de ocorrência de incêndios utilizando dados meteorológicos públicos, contribuindo para o desenvolvimento de ferramentas de apoio ao monitoramento ambiental e ao planejamento de ações preventivas.

REFERÊNCIAS

BDQUEIMADAS - INPE. **Banco de Dados de Queimadas**. Disponível em: https://terrabrasilis.dpi.inpe.br/queimadas/bdqueimadas/. Acesso em: 30 ago. 2025.

DUARTE, Miqueias Lima. **Previsão da suscetibilidade a incêndios e queimadas utilizando um modelo baseado em inteligência artificial e sistema de inferência fuzzy**. Sorocaba: UNESP, 2022. Disponível em: http://hdl.handle.net/11449/217103

DUTRA, Gabriel. **Previsão de eventos emergenciais utilizando séries temporais e aprendizado de máquina**. Florianópolis, 2024. Dissertação (Mestrado) Universidade Federal de Santa Catarina. Disponível em: https://repositorio.ufsc.br/handle/123456789/262178. Acesso em: 30 jun. 2026.

EMBRAPA. **Bioma Caatinga**: introdução. 2021. Disponível em: https://www.embrapa.br/agencia-de-informacao-tecnologica/tematicas/bioma-caatinga/introducao. Acesso em: 5 jan. 2026.

HANAMOTO, Juniti Hitochi Afonço. ***Machine learning approaches for fire risk classification using remote sensing and meteorological data***. 2025. Dissertação (Mestrado) - Universidade de São Paulo, São Paulo, 2025. Disponível em: https://teses.usp.br/teses/disponiveis/3/3152/tde-09042026-152014/. Acesso em: 14 jul. 2026.

ICMBIO. **Plano de Manejo Integrado do Fogo** - Parque Nacional da Serra da Capivara. 2023. Disponível em: https://www.gov.br/icmbio/pt-br/centrais-de-conteudo/publicacoes/planos-de-manejo-integrado-do-fogo/. Acesso em: 15 jun. 2026.

INSTITUTO NACIONAL DE METEOROLOGIA (INMET). **Banco de Dados Meteorológicos para Ensino e Pesquisa (BDMEP)**. Brasília: INMET. Disponível em: https://bdmep.inmet.gov.br/. Acesso em: 5 jan. 2026.

INSTITUTO NACIONAL DO SEMIÁRIDO (INSA). **Dia da Caatinga**: o papel da pesquisa e das ações em rede na proteção da região semiárida com maior biodiversidade do mundo. Campina Grande: INSA, 2025. Disponível em: https://www.gov.br/insa/pt-br/assuntos/noticias/dia-da-caatinga-o-papel-da-pesquisa-e-das-acoes-em-rede-na-protecao-da-regiao-semiarida-com-maior-biodiversidade-do-mundo. Acesso em: 12 jul. 2026.

INSTITUTO NACIONAL DE PESQUISAS ESPACIAIS (INPE). **Plano de Prevenção e Controle do Desmatamento e das Queimadas na Caatinga (PPCaatinga 2024)**. Brasília: Ministério do Meio Ambiente e Mudança do Clima, 2024. Disponível em: https://www.gov.br/mma/pt-br/assuntos/controle-ao-desmatamento-queimadas-e-ordenamento-ambiental-territorial/controle-do-desmatamento-1/ppcaatinga/ppcaatinga2024_20_12.pdf. Acesso em: 16 jul. 2026.

KE, Guolin et al. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. ***Advances in Neural Information Processing Systems***, v. 30, 2017.

MAPBIOMAS. **Área queimada no Brasil em janeiro aumentou 3,5 vezes em relação a 2023**. 2024. Disponível em: https://brasil.mapbiomas.org/2024/02/29/area-queimada-no-brasil-em-janeiro-aumentou-35-vezes-em-relacao-a-2023/. Acesso em: 16 nov. 2025.

MOURA, Lucas. **O papel da inteligência artificial no monitoramento e prevenção de desastres ambientais**. São Paulo, 2025. Trabalho de Conclusão de Curso (Graduação) - Faculdade de Tecnologia de São Paulo. Disponível em: https://ric.cps.sp.gov.br/handle/123456789/35420. Acesso em: 30 jun. 2026.

NEIRA, Lucas et al. Utilização de machine learning para identificação de focos de queimadas por imagens. **Revista Contemporânea**, v. 4, n. 12, 2024. DOI: https://doi.org/10.56083/RCV4N12-124.