UNIVERSIDADE FEDERAL DE SÃO CARLOS
CENTRO DE CIÊNCIAS EXATAS E DE TECNOLOGIA - CCET
DEPARTAMENTO DE ENGENHARIA ELÉTRICA - DEE

JÚLIO YAMAMOTO

Modelos preditivos aplicados ao risco de incêndios no Cerrado: desempenho e importância de variáveis ambientais

São Carlos - SP
2025

JÚLIO YAMAMOTO

Modelos preditivos aplicados ao risco de incêndios no Cerrado: desempenho e importância de variáveis ambientais

Trabalho de conclusão de curso apresentado ao Departamento de Engenharia Elétrica da Universidade Federal de São Carlos, para obtenção do título de bacharel em Engenharia Elétrica.

Orientador: Prof. Dr. Celso Aparecido de França.

São Carlos - SP
2025

Resumo

A recorrência de queimadas no Cerrado brasileiro compromete severamente a fauna, a flora e a biodiversidade do bioma. Entre 2001 e 2019, estudos apontaram redução na atividade fotossintética das folhas das árvores da região, atribuída à frequência desses eventos. Neste contexto, o presente trabalho teve como objetivo avaliar o impacto de variáveis ambientais e meteorológicas na definição do risco de fogo. Para isso, foram aplicados três modelos de aprendizado de máquina com abordagens distintas (*Random Forest*, *XGBoost* e *MLPRegressor*) com o intuito de replicar o índice de risco de fogo disponibilizado pelo INPE, utilizando um conjunto reduzido de cinco variáveis: temperatura do ar, precipitação acumulada, número de dias consecutivos sem chuva, umidade relativa e concentração atmosférica de fumaça. Foram analisados mais de 600 arquivos diários cobrindo o período entre agosto de 2023 e maio de 2025. Após a fase de treinamento e avaliação dos modelos, utilizando técnicas de validação cruzada e métricas de regressão, o modelo baseado em *Random Forest* obteve o melhor desempenho preditivo, com coeficiente de determinação (R^2) final de 0,8166. A partir da definição deste modelo, realizaram-se análises de importância das variáveis (por permutação e por correlação), análises residuais, testes de sensibilidade e comparações espaciais das predições. Verificou-se que três variáveis — umidade relativa, número de dias sem chuva e precipitação — foram responsáveis por cerca de 90% da capacidade explicativa do modelo. Os testes de sensibilidade confirmaram essa influência, com o aumento do número de dias sem chuva e a redução da umidade e da precipitação resultando em elevação do risco previsto. Adicionalmente, as análises residuais demonstraram erros centrados em torno de zero, com baixa variabilidade e ausência de viés sistemático, enquanto os mapas comparativos mensais revelaram boa aderência espacial entre os valores previstos e os observados. Apesar dos resultados expressivos, o trabalho apresenta limitações importantes: o número reduzido de variáveis, a ausência de dados sobre cobertura vegetal, vento, pressão atmosférica, e concentração de gases na atmosfera. Além disso, a incompatibilidade entre a escala temporal do índice de risco (diária) e das áreas queimadas (mensal) dificultam avaliações mais precisas no estabelecimento de correlações entre o índice e seu impacto. Recomenda-se, em trabalhos futuros, a incorporação de variáveis adicionais e dados complementares que permitam aumentar a robustez do modelo e sua capacidade de generalização.

Palavras-chave: risco de fogo; incêndios florestais; aprendizado de máquina; Cerrado; importância de variáveis.

Abstract

The recurrence of wildfires in the Brazilian Cerrado severely threatens the region's fauna, flora, and overall biodiversity. Between 2001 and 2019, studies reported a decline in the photosynthetic activity of tree foliage in the area, attributed to the frequent occurrence of these fires. In this context, the present study aimed to assess the impact of environmental and meteorological variables on fire risk estimation. Three machine learning models with distinct approaches were applied (Random Forest, XGBoost, and MLPRegressor) to replicate the fire risk index provided by INPE, using a reduced set of five variables: air temperature, accumulated precipitation, number of consecutive dry days, relative humidity, and atmospheric smoke concentration. Over 600 daily files covering the period from August 2023 to May 2025 were analyzed. After training and evaluating the models using cross-validation and regression metrics, the Random Forest model achieved the best predictive performance, with a final coefficient of determination (R^2) of 0.8166. Based on this model, further analyses were conducted, including permutation and correlation-based feature importance, residual diagnostics, sensitivity tests, and spatial comparisons of predictions. Three variables—relative humidity, number of dry days, and precipitation—accounted for approximately 90% of the model's predictive power. The sensitivity analysis confirmed this influence, as increased dry days and reduced humidity and precipitation led to higher predicted fire risk. Residual analysis showed errors centered around zero, with low variability and no systematic bias, while monthly spatial maps demonstrated strong alignment between predicted and observed values. Despite promising results, the study has notable limitations: the reduced number of variables, the absence of data on vegetation cover, wind, atmospheric pressure, and gas concentrations. Moreover, the mismatch between the temporal resolution of fire risk data (daily) and burned area data (monthly) hinders precise correlation between predicted risk and actual impact. Future work should incorporate additional environmental variables and complementary datasets to enhance model robustness and generalizability.

Keywords: risk of fire; wildfires; machine learning; Cerrado; feature importance.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

DT	árvores de decisão, do inglês *decision tree*
IDW	ponderação inversa pela distância, do inglês *inverse distance weighting*
INPE	Instituto Nacional de Pesquisas Espaciais
KNN	k vizinhos mais próximos, do inglês *k-nearest neighbors*
MAE	erro médio absoluto, do inglês *mean absolute error*
ML	aprendizado de máquina, do inglês *machine learning*
MLP	perceptron multicamadas, do inglês *Multilayer Perceptron*
MSE	erro quadrático médio, do inglês *mean square error*
RF	floresta aleatória, do inglês *random forest*
RNA	rede neural artificial
RMSE	raiz do erro quadrático médio, do inglês *root mean square error*
SMAPE	erro percentual absoluto médio simétrico, do inglês *symmetric mean absolute percentage error*
SVM	máquina de vetores de suporte, do inglês *support vector machine*

Sumário

1 Introdução

O presente trabalho aborda a aplicação de técnicas de aprendizado de máquina na estimativa do risco de incêndios florestais, com foco no bioma Cerrado, bem como a análise do impacto das variáveis ambientais e climatológicas nesse contexto. A motivação para o estudo decorre da crescente recorrência de queimadas na região, fenômeno que ameaça a biodiversidade local. Este capítulo apresenta a contextualização do problema, os objetivos da pesquisa, a justificativa para sua realização e a organização geral do trabalho.

1.1 Contextualização

O bioma do Cerrado brasileiro é um ecossistema singular e particular na América do Sul. Ocupando uma vasta área territorial que representa aproximadamente um quarto do Brasil, o Cerrado é possuidor de uma extensa e rica biodiversidade na região. Esta, que, por sua vez, é atualizada a cada ano através da descoberta de novas espécies, evidenciando ainda mais seu enorme potencial biológico que ainda não foi totalmente explorado (COLLI; VIEIRA; DIANESE, 2020).

Reconhecendo seu valor ecológico, torna-se fundamental a compreensão e o entendimento de atividades que ponham a integridade das florestas em risco. Enquanto campanhas de conscientização possuam o maior foco recaído sobre o desmatamento e obtenham relativo sucesso, outras formas de degradação florestal também devem ser estudadas. Em 2015, por exemplo, as taxas de desmatamento na Amazônia caíram 66% quando comparadas à média observada entre 1988 e 2004. Apesar disso, durante o mesmo espectro temporal, a incidência de queimadas na região aumentou 36% (ARAGãO et al., 2018). No Cerrado, Lemos (2024) destaca o aumento de 31% no número de focos de queimadas na região durante 2024, se comparado ao período homólogo.

No mesmo ano, a Confederação Nacional de Municípios (CNDM) informou que as queimadas ocorridas no Brasil provocaram prejuízos econômicos de aproximadamente R$ 1,3 bilhão, conforme apontado pela CNN (2024). As queimadas afetam diretamente as características do ecossistema, empobrecendo a qualidade da água e do solo, além de prejudicar o hábitat da fauna local. A longo prazo, tais ações resultam numa redução da biodiversidade local, afetando processos ecológicos como o ciclo de carbono, o clima e a capacidade regenerativa do solo em relação à sua degradação (HARPER et al., 2018; PELLEGRINI et al., 2018). Além disso, Oliveira et al. (2022) evidenciam como a presença de queimadas e seu risco de disseminação podem ameaçar as estruturas, residências e

plantações próximas, colocando a população local em estado de alerta. A emissão de gases poluentes durante o processo de queima também deve ser considerada, devido ao impacto direto na saúde pública.

A análise quantitativa de focos de incêndio não é a única métrica considerada na avaliação do impacto das queimadas. Fidelis et al. (2018) chama a atenção, principalmente, para a área queimada pelo fogo. Em 2017, enquanto o número de focos ativos no Cerrado permaneceu similar em relação à média observada em anos anteriores, a média anual da área queimada cresceu significativamente. Os intervalos de tempo de 2011–2014 e 2015–2017 registraram um aumento na área queimada de 100% no Cerrado, respectivamente. O autor atrela esse fato, principalmente, à elevada ocorrência de mega incêndios na região, i.e., queimadas caracterizadas não só pela sua extensão, mas, principalmente, pelo impacto causado e pela dificuldade em extingui-las.

Convém salientar que a inibição total da presença de fogo no Cerrado não é a forma ideal de lidar com o problema. Durigan (2020) alerta que, apesar da publicidade governamental alertar a população sobre o risco das queimadas, sua completa supressão não é a forma indicada de lidar com esse fenômeno. As queimadas já faziam parte do bioma antes da presença humana na região, de tal forma que sua vegetação apresenta sinais de adaptabilidade e depende do fogo para manter sua reprodução e sobrevivência. A discriminação contra o uso de fogo em reservas, por exemplo, causou quedas na biodiversidade da região em níveis territoriais, de espécie e de população. Assim, o objetivo deste trabalho não é eliminar todo e qualquer foco ativo, mas sim contribuir para o aprimoramento das estratégias de monitoramento e combate ao fogo, com base em critérios técnicos e contextuais.

Diante desse cenário, este trabalho propõe o uso de modelos baseados em técnicas de aprendizado de máquina para prever o risco de incêndios florestais no Cerrado, a partir das medições de variáveis meteorológicas e atmosféricas. A proposta visa não apenas antecipar áreas com maior propensão à ocorrência de queimadas, mas também identificar as variáveis mais relevantes na formação desses eventos. Com isso, busca-se oferecer subsídios para a atuação mais eficiente dos órgãos responsáveis, otimizando recursos e permitindo intervenções preventivas que reduzam o impacto dos incêndios sem comprometer o funcionamento ecológico do bioma.

1.2 Objetivos

Para orientar o desenvolvimento deste estudo, define-se inicialmente um objetivo geral, seguido por objetivos específicos que desdobram as ações a serem realizadas ao longo do trabalho.

1.2.1 Objetivo Geral

O objetivo deste trabalho consiste em, a partir de dados meteorológicos e atmosféricos disponíveis ao público através do Programa Queimadas do Instituto Nacional de Pesquisas Espaciais (INPE), desenvolver modelos preditivos capazes de replicar o índice de risco de fogo disponibilizado pelo programa, com foco especial na região do Cerrado brasileiro. A partir disso, pretende-se avaliar o impacto de cada uma das variáveis, investigar a viabilidade da estimação do índice a partir de um conjunto reduzido de entradas e analisar a sensibilidade do modelo mediante exclusão de diferentes parâmetros. Dessa forma, busca-se aprofundar a compreensão sobre os fatores que influenciam a construção do índice e oferecer subsídios para futuras aplicações em sistemas de alerta e planejamento de ações preventivas.

1.2.2 Objetivos Específicos

- Coletar e organizar dados ambientais relevantes — como temperatura média, precipitação acumulada, número de dias consecutivos sem chuva, umidade relativa do ar e concentração atmosférica de fumaça — em recortes enquadrados ao Cerrado brasileiro.

- Treinar modelos supervisionados de aprendizado de máquina, do inglês *machine learning* (ML), estabelecendo como variável-alvo o índice de risco de fogo disponibilizado publicamente pelo INPE.

- Avaliar o desempenho dos modelos treinados com o auxílio de métricas apropriadas a análise de regressões, como o coeficiente de determinação (R^2), MAE, MSE, RMSE e o SMAPE, identificando a capacidade do modelo treinado em capturar corretamente o comportamento do valor do índice de risco de fogo do INPE.

- Realizar análise da importância das variáveis (*feature importance*) para identificar quais fatores ambientais exercem maior influência na estimativa do índice de risco.

- Realizar análise residual para investigar padrões de erro nas predições do modelo, identificando possíveis viéses ou inconsistências nas predições realizadas.

- Avaliar a sensibilidade das predições às variáveis de entrada por meio de testes de dependência parcial e análise SHAP, visando avaliar a robustez e o comportamento do modelo em relação a diferentes entradas.

- Produzir comparações espaciais gráficas com base nos resultados obtidos, com visualizações do índice do risco de fogo oficial, previsto e da diferença observada entre

ambos, além de gráficos de dispersão das variáveis de entrada, distribuição dos erros e correlação entre variáveis.

- Contextualizar os achados com base na literatura científica existente sobre queimadas no Cerrado, discutindo como os fatores ambientais modelados se alinham (ou não) às conclusões já estabelecidas por estudos anteriores.

1.3 Justificativa

É relevante destacar que o foco das medidas contra as queimadas não pode ser restringido somente ao ponto de ignição ou na constatação de focos ativos na região. As origens de um incêndio podem ser naturais, antrópicas ou desconhecidas, e por isso, torna-se imprescindível a compreensão das tendências favoráveis à sua ocorrência e disseminação, permitindo que órgãos responsáveis atuem de forma antecipada e estratégica.

A situação é tão crítica que, apesar da adoção de políticas governamentais voltadas à redução de danos por incêndios florestais em 2017, cortes no orçamento entre 2019 e 2021 causaram uma paralisação no cumprimento dessas regras. Consequentemente, uma onda de queimadas associadas à alta do desmatamento causou a devastação de grandes áreas tanto na Amazônia quanto no Cerrado (OLIVEIRA et al., 2021).

Nesse contexto, a utilização de técnicas de ML mostra-se promissora por sua capacidade de lidar com grandes volumes de dados ambientais e identificar padrões não lineares com alto grau de precisão. Tais ferramentas podem atuar como mecanismos de suporte, fornecendo subsídios técnicos em situações em que a atuação humana é limitada por tempo, recursos ou escala.

A escolha do tema também tem origem na aplicação de conhecimentos técnicos adquiridos na formação do curso em um problema concreto de relevância social e ambiental. O foco no Cerrado — bioma frequentemente negligenciado em comparação à Amazônia, mas igualmente vulnerável ao fogo — busca contribuir para a discussão sobre o uso de ferramentas computacionais no apoio ao monitoramento e à formulação de políticas públicas mais eficazes.

1.4 Estrutura do Trabalho

O trabalho está dividido em seis capítulos. Este capítulo inicial apresenta a contextualização e a justificativa do trabalho. No segundo capítulo, é apresentada a fundamentação teórica sobre o tema, baseada na revisão bibliográfica realizada. Esta, por sua vez, é apresentada no terceiro capítulo. O quarto capítulo descreve o desenvolvimento e a

implementação do trabalho. No quinto capítulo, são relatados e discutidos os resultados obtidos. Por fim, o sexto capítulo apresenta as considerações finais.

2 Fundamentação Teórica

Os incêndios florestais no Cerrado têm provocado impactos severos sobre sua vegetação nativa, biodiversidade e funcionamento ecológico. Entre 1985 e 2023, o fogo consumiu 88 milhões de hectares do Cerrado — extensão territorial comparável à Venezuela. Isso significa que, em média, o bioma teve 9,5 milhões de hectares queimados anualmente, registrando índices maiores que os da Amazônia, que registrou a queima de 7,1 milhões de hectares anualmente (GUARALDO, 2024). Com 34% de sua extensão territorial afetada pelo fogo, o Cerrado é o segundo bioma mais atingido do país, atrás somente do Pantanal, com 45%. Apesar da vegetação local apresentar uma série de adaptações morfofisiológicas ao fogo, observou-se na região uma redução de 19% na atividade fotossintética das folhas entre os anos 2001 e 2019, indicando um declínio progressivo na capacidade regenerativa da cobertura vegetal após sucessivos eventos de queimada (OLIVEIRA et al., 2022).

2.1 Variáveis Meteorológicas, Climatológicas e Ambientais

A ocorrência e propagação dos incêndios florestais estão diretamente relacionadas com as condições atmosféricas e características do ambiente. Neste trabalho, foram selecionadas variáveis meteorológicas, climatológicas e ambientais cujos dados estivessem disponíveis publicamente com alto grau de confiabilidade, e que, simultaneamente, exercessem influência significativa sobre o risco de fogo, baseando-se na literatura.

As variáveis consideradas possuem valores obtidos a partir da observação da média diária. Considerou-se a temperatura do ar (K), a umidade relativa do ar (%), a precipitação acumulada (mm dia^{-1}), o número de dias consecutivos sem chuva e a concentração atmosférica de fumaça (μg/m^3).

A temperatura do ar está frequentemente associada à inflamabilidade da matéria orgânica, influenciando diretamente a severidade e a extensão dos focos de incêndio. Já a umidade relativa do ar influencia o grau de secura da matéria orgânica, isto é, atuando diretamente como um fator de medida de combustibilidade da vegetação, e ditando o comportamento da queimada (WASSERMAN; MUELLER, 2023). A precipitação acumulada e o número de dias consecutivos sem chuva são variáveis complementares: enquanto a primeira aponta para eventos recentes de chuva, a segunda indica o grau de estiagem prolongada. Ambas possibilitam a categorização adequada da combustibilidade do ambiente, isto é, seu potencial de queima (SETZER; SISMANOGLU; SANTOS, 2019). A concentração atmosférica da fumaça, por sua vez, pode representar tanto a ocorrência recente

de incêndios quanto a permanência de material particulado no ar. Aqui, é utilizada para auxiliar no estabelecimento de uma correlação com o índice de risco de fogo calculado.

Os dados utilizados estão disponíveis de forma aberta e gratuita ao público, obtidos diretamente através do Programa Queimadas do INPE. Considerou-se toda a amostragem temporal em que estavam disponíveis, simultaneamente, valores válidos para todas as variáveis desejadas. Por isso, o estudo incluiu todos os dados diários entre agosto de 2023 e maio de 2025.

As informações são derivadas de sensores ópticos operando na faixa termal-média de 4μm, instalados em dez satélites distintos. Esses satélites incluem plataformas polares — como as AVHRR/3 dos NOAA-18 e 19, METOP-B e C, MODIS dos satélites TERRA e AQUA, e VIIRS do NPP-Suomi e NOAA-20 — e geoestacionárias, como o GOES-16 e o MSG-3. Posteriormente, as imagens são processadas na Divisão de Geração de Imagens (DGI) e na Divisão de Satélites e Sistemas Ambientais (DSA) (PROGRAMA QUEIMADAS DO INPE, 2025). Esses satélites realizam suas passagens constantemente nos mesmos horários sobre um mesmo ponto da Terra, mantendo um padrão estável de coleta ao longo do tempo.

Os dados de temperatura, umidade relativa, número de dias sem chuva, precipitação acumulada e o índice de risco de fogo desenvolvido pelo INPE são disponibilizados em arquivos diários em formato *.nc*. Cada arquivo associa os valores das variáveis a coordenadas específicas de latitude e longitude, cobrindo toda a extensão do território brasileiro, e com resolução espacial de 1 km^2 por célula. A concentração atmosférica de fumaça, por sua vez, é disponibilizada em formato *.tiff*, e segue ao mesmo princípio de organização. O tratamento e manuseio de ambos é descrito com detalhes no Capítulo 4.

2.2 Técnicas de Interpolação de Dados

O tratamento e manipulação de dados envolve, por vezes, a necessidade de estimar valores em regiões onde não há observações diretas, baseando-se somente na informação de pontos vizinhos. Esse procedimento se torna necessário quando o conjunto de amostras é escasso, ou quando os dados disponíveis não compartilham uma mesma grade de referência espacial. Serão apresentadas técnicas adequadas ao escopo deste estudo, cuja necessidade é demonstrada com mais detalhes no Capítulo 4. A escolha dos métodos baseia-se em diferentes densidades amostrais, nas características dos dados ou limitações computacionais oriundas do processamento dos dados.

2.2.1 Interpolação Linear

Burden e Faires (2011) descrevem a interpolação linear como a construção de uma função que estima um valor intermediário baseado em dois pontos adjacentes cujos valores são conhecidos. Esse método é um caso particular da interpolação polinomial, em que a função interpoladora é de grau 1. Dado dois pontos conhecidos $A(x_a, y_a)$ e $B(x_b, y_b)$, é possível definir a função interpoladora $f(x)$ entre esses pontos para um ponto (x, y) por:

$$f(x) = y = y_a + (y_b - y_a)\frac{(x - x_a)}{(x_b - x_a)} \tag{2.1}$$

Em um contexto bidimensional — como o tratamento de dados geográficos, por exemplo — a interpolação linear pode ser estendida pela decomposição do espaço em triângulos, por meio da triangulação de Delaunay. Essa técnica divide um conjunto de pontos em triângulos não sobrepostos, buscando maximizar os menores ângulos internos. A partir dessa malha triangular, os valores em qualquer ponto interno são estimados por uma função linear baseada nos valores dos vértices. Essa abordagem proporciona uma transição suave entre os pontos interpolados, o que a torna adequada para aplicações com distribuição espacial irregular.

2.2.2 Interpolação pelo Inverso da Distância

O método da ponderação inversa pela distância, do inglês *inverse distance weighting* (IDW) é uma das técnicas de interpolação mais amplamente utilizadas por geocientistas e pesquisadores ambientais. Sua premissa baseia-se na suposição de que os valores de uma variável em pontos distintos do espaço são correlacionados, mas sua influência decresce com o aumento da distância entre os pontos. O método é especialmente útil em aplicações envolvendo grandes volumes de dados, por permitir a limitação da interpolação a um subconjunto de vizinhos próximos utilizando a técnica dos k vizinhos mais próximos, do inglês *k-nearest neighbors* (KNN), reduzindo o custo computacional sem comprometer significativamente a qualidade da estimativa. Conforme apontado por Lu e Wong (2008), a interpolação em um ponto desejado S_0 é dada por:

$$\hat{y}(S_0) = \sum_{i=1}^{n} \lambda_i y(S_i) \tag{2.2}$$

em que $\hat{y}(S_0)$ representa o valor interpolado no ponto S_0, com base nos valores conhecidos $y(S_i)$ de n pontos vizinhos. Por outro lado, λ_i representa os pesos atribuídos a cada ponto, cuja soma deve ser igual a 1. Isso significa que $\hat{y}(S_0)$ representa uma combinação linear ponderada dos valores conhecidos. Os pesos são definidos por:

$$\lambda_i = \frac{d_{0i}^{-\alpha}}{\sum_{i=1}^{n} d_{0i}^{-\alpha}} \tag{2.3}$$

onde d_{0i} é a distância entre o ponto de interesse S_0 e S_i. α é o parâmetro de potência que controla a influência da distância. Valores maiores α aumentam a influência dos pontos mais próximos, enquanto valores menores atenuam essa diferença, permitindo maior influência de pontos mais distantes.

2.3 Conceitos de Aprendizado de Máquina

A primeira utilização do termo "aprendizado de máquina" (ML) foi feita por Samuel (1959). Na prática, consiste no fornecimento de dados ao computador e na orientação de suas decisões, de forma que o mesmo execute determinadas ações e obtenha respostas desejadas. Mitchell (1997) descreve o termo como sendo uma área focada no desenvolvimento de algoritmos que buscam aprimorar a execução de certas tarefas com o intermédio da análise de observações, tendo consigo uma métrica de desempenho como referência. Na literatura moderna, o tópico pode ser categorizado a partir de três vertentes principais, descritas abaixo:

2.3.1 Aprendizado Supervisionado

Os modelos baseados em aprendizado supervisionado possuem uma abordagem de treinamento de modelos onde os dados de entrada e de saída são relacionados, ou seja, cada entrada possui uma saída esperada associada. O objetivo principal do modelo é corretamente associar entradas e saídas esperadas, minimizando o erro de previsão em novos dados (KOTSIANTIS; ZAHARAKIS; PINTELAS, 2007). Podem ser subdivididos em problemas de:

- **Classificação**: Quando o objetivo é atribuir entradas a categorias previamente definidas, com base em dados rotulados. Após o treinamento, o modelo consegue classificar dados não classificados. Um exemplo são modelos de reconhecimento de imagens.
- **Regressão**: Quando o objetivo é voltado à previsão de valores contínuos. Um exemplo comum é a previsão de preços, ou de índices, como no caso deste trabalho.

2.3.2 Aprendizado Não-Supervisionado

Modelos de aprendizado não supervisionado buscam identificar padrões a partir da entrada de um conjunto de dados sem rótulos. São aplicados em tarefas voltadas à detecção

de anomalias ou agrupamentos (*clusterização*), onde não se atribui uma resposta correta anteriormente. Ao invés de prever saídas, os modelos buscam compreender a relação entre os dados e como se organizam. É empregado, por exemplo, em análises de sentimento de textos — área de estudo voltada à identificação e classificação de opiniões e declarações, visando determinar a polaridade associada a um determinado tópico (MITTAL; PATIDAR, 2019).

2.3.3 Aprendizado por Transferência

O aprendizado por transferência baseia-se na reutilização de um modelo pré-treinado aplicado em um novo problema, geralmente com menos dados. A lógica por trás é realizar o treinamento do modelo com um grande conjunto de dados e aplicá-lo em tarefas onde a disponibilidade de dados é escassa. Dessa forma, é possível acelerar o aprendizado ou aumentar a acurácia do modelo, enquanto se economiza gastos envolvidos na reconstrução total do modelo. Hoje em dia, observa-se a aplicação desses modelos em robôs autônomos (RIS-ALA, 2023).

2.4 Modelos de Aprendizado Supervisionado Aplicados à Regressão

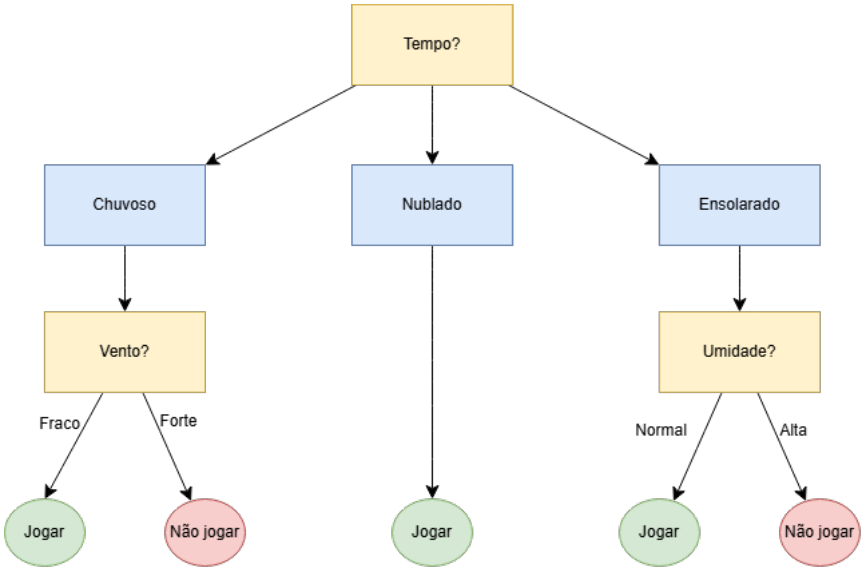
Considerando o contexto do tema inicial deste trabalho (predição do índice de risco de fogo fornecido pelo INPE), propõe-se aplicar modelos de ML supervisionados com foco na resolução de um problema de regressão. Serão apresentados nas subseções a seguir três modelos de regressão distintos, cuja seleção foi baseada na recorrência em aplicação na literatura e na complementaridade entre suas estruturas algorítmicas: *Random Forest*, *XGBoost* e *MLPRegressor* (CHEN; GUESTRIN, 2016; BREIMAN, 2001; GOODFELLOW; BENGIO; COURVILLE, 2016).

2.4.1 Árvores de Decisão

As árvores de decisão, do inglês *decision tree* (DT) é o modelo base de ambos *Random Forest* e *XGBoost*. Suas decisões são estruturadas a partir de dados de entrada por meio de uma representação hierárquica em forma de árvore. É composta inicialmente por um nó raiz, que representa o ponto de partida das tomadas de decisão. A partir dela, cada nó interno da árvore representa uma decisão baseada em uma variável de entrada (atributo), e cada aresta que liga os nós representa o resultado dessa decisão — uma condição verdadeira ou falsa, por exemplo. As folhas (também chamadas de nós terminais) representam a saída final do modelo, que pode ser um valor contínuo (em modelos de regressão) ou classe (em modelos de classificação) (MAIMON; ROKACH, 2010). A Figura

1 apresenta uma árvore de decisão que exemplifica o processo de escolha sobre jogar ou não tênis.

Figura 1 – Exemplo de Árvore de Decisão sobre jogar ou não tênis



Fonte: Autoria própria

O objetivo primário de uma DT é, a partir de uma amostra de N dados com k atributos, obter a melhor classificação possível da amostra, mantendo o menor erro de generalização — ou seja, garantir um bom desempenho preditivo quando exposta a dados não vistos. Um dos desafios nesse processo é evitar o sobreajuste (*overfitting*), que ocorre quando as DTs de N dados com k atributos geram exatamente N folhas, i.e., cada exemplo da amostra é perfeitamente memorizado. A classificação torna-se, então, incapaz de generalizar e torna o modelo ineficaz em novas aplicações. Por outro lado, um modelo excessivamente simplificado pode sofrer de subajuste (*underfitting*), que também deve ser evitado. Neste, o modelo falha ao capturar padrões relevantes de dados e apresenta, por consequência, desempenho insatisfatório em amostras de treino e de teste (MAIMON; ROKACH, 2010).

2.4.2 Random Forest

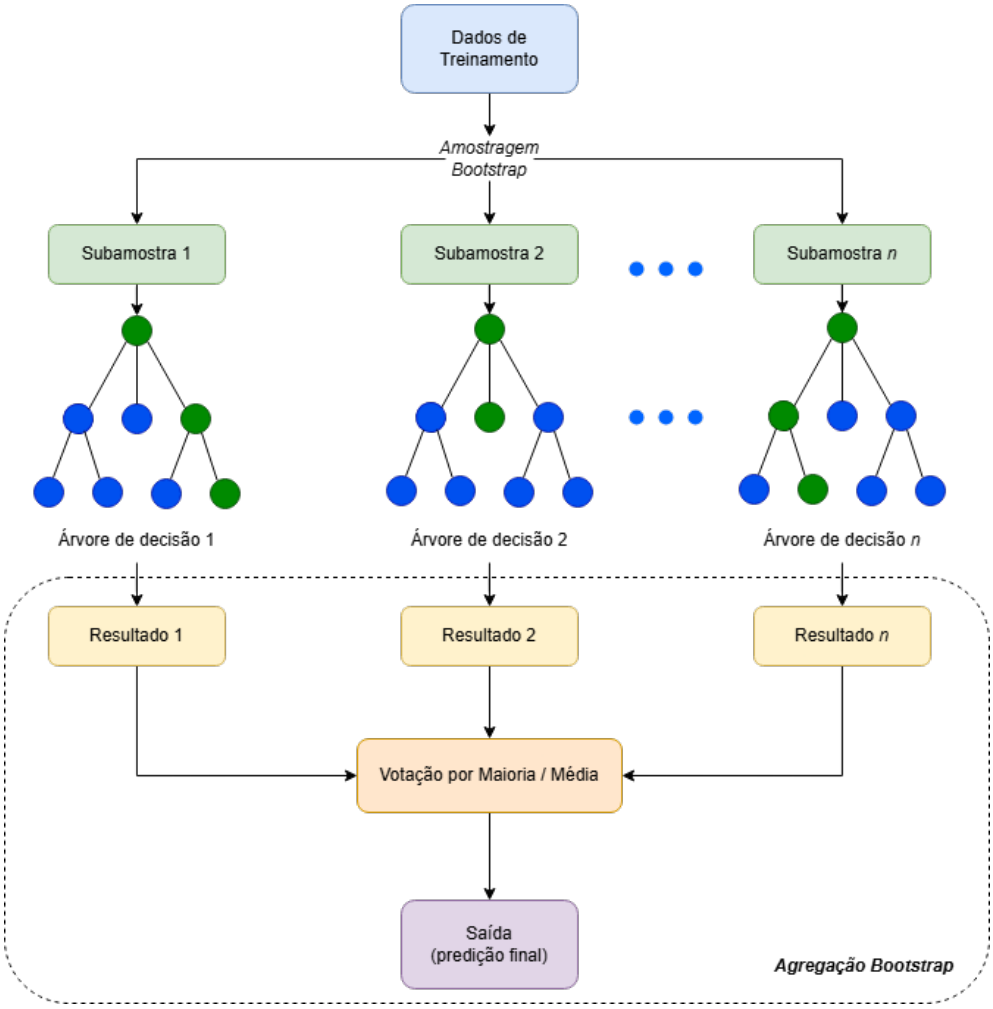
A floresta aleatória, do inglês *random forest* (RF) é um algoritmo baseado no conceito de DT, cuja proposta é construir uma "floresta" de árvores de decisão independentes, cada uma delas otimizada a partir de subconjuntos aleatórios dos dados, visando melhorar a capacidade de generalização do modelo (reduzindo o risco de *overfitting*). Seu diferen-

cial está na aplicação conjunta das técnicas de *bootstrap* e *aggregation*, que juntas são denominadas *bagging*.

A técnica de *bootstrap* refere-se à criação de várias subamostras a partir de um conjunto inicial de D amostras e M atributos, utilizando seleção com reposição, ou seja, cada amostra do conjunto original pode ser escolhida mais de uma vez para compor uma subamostra. O processo resulta em subconjuntos aleatórios contendo $d < D$ amostras e $m < M$ atributos. Já a *aggregation* consiste na combinação dos resultados individuais dessas subamostras para a obtenção de uma previsão agregada, por média (no caso de regressão) ou por votação majoritária (no caso de classificação), reduzindo a variância do modelo final. Isso acaba gerando um modelo mais estável e menos suscetível a distorções (MAIMON; ROKACH, 2010).

Segundo Ayyadevara (2018), o funcionamento da RF pode ser resumido em seis etapas: 1) seleção de um subconjunto aleatório de atributos; 2) geração de uma amostra com reposição (*bootstrap*) do conjunto de dados original; 3) construção de uma árvore de decisão com base no subconjunto da amostra; 4) repetição dos passos anteriores n vezes, sendo n o número total de árvores; 5) geração de previsões independentes por todas as árvores no conjunto de dados; e 6) geração da saída final a partir da combinação das previsões por média (regressão) ou maioria (classificação). A Figura 2 apresenta um diagrama de blocos simplificado que ilustra o funcionamento de uma RF, destacando as etapas de amostragem e agregação *bootstrap*.

Figura 2 – Diagrama explicativo do algoritmo *Random Forest*



Fonte: Autoria própria

Hastie, Tibshirani e Friedman (2009) também destacam uma característica importante da RF, as *out-of-bag* (OOB) *samples*. Os autores a descrevem da seguinte forma:

> *For each observation* $z_i = (x_i, y_i)$*, construct its random forest predictor by averaging only those trees corresponding to bootstrap samples in which* z_i *did not appear.* (HASTIE; TIBSHIRANI; FRIEDMAN, 2009)

Em outras palavras, o conceito refere-se ao uso de observações que não foram selecionadas nas amostras *bootstrap* utilizadas para treinar uma determinada árvore. As amostras excluídas (chamadas de *out-of-bag*) podem ser utilizadas para testar a árvore construída sem a necessidade de um conjunto de validação externo, permitindo uma estimativa interna do erro de generalização do modelo. O método funciona de forma análoga à validação cruzada *k-fold* — discutida posteriormente na seção de Métricas de Avaliação

— mas é mais eficiente computacionalmente, por aproveitar diretamente o processo de *bagging* durante o treinamento.

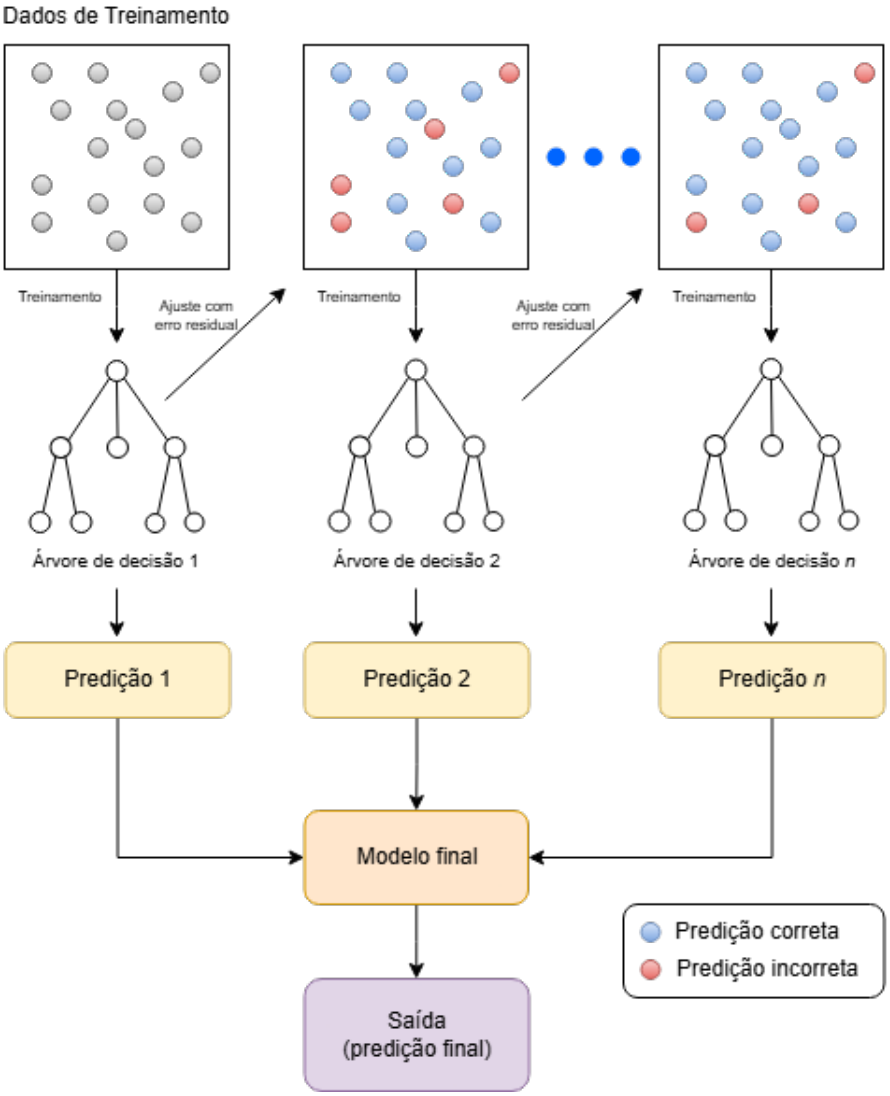
2.4.3 XGBoost

O XGBoost (*Extreme Gradient Boosting*), assim como o RF, também é um algoritmo baseado no conceito de DT, mas com uma grande diferença na abordagem: ao invés de construir árvores de forma independente e agregar seus resultados — com funcionamento baseado na técnica de *bagging*, o modelo prioriza o *boosting*, onde as árvores são criadas de maneira sequencial. A ideia principal é corrigir iterativamente os erros cometidos por modelos anteriores e, para isso, atribui-se maior peso às observações mal previstas nas iterações seguintes. Dessa forma, cada nova árvore busca reduzir o erro residual das árvores anteriores (HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

A implementação proposta por Chen e Guestrin (2016) aprimora o método tradicional de *gradient boosting* por introduzir otimizações como a paralelização do processo de construção de árvores, técnicas de controle de sobreajuste (*overfitting*) e o uso de *subsampling* de colunas e linhas. Se destaca de outros métodos por usar uma função objetivo regularizada (*regularized objective function*), que adiciona um termo de regulação à função de perda tradicional, que considera tanto o número de folhas quanto a magnitude de seus pesos, promovendo modelos simplificados e com menor risco de sobreajuste. O método *AdaBoost*, por exemplo, ao atribuir pesos maiores a predições incorretas, torna-se sensível à influência de *outliers* durante o ajuste iterativo do modelo.

Como resultado, o *XGBoost* é capaz de se ajustar melhor ao ruído dos dados, simplificando-os e melhorando sua capacidade de generalização. O algoritmo se popularizou por sua eficácia em cenários com grandes volumes de dados e na presença de variáveis ruidosas. Uma limitação, no entanto, é sua sensibilidade a hiperparâmetros, que demanda ajuste fino durante o desenvolvimento do modelo (CHEN; GUESTRIN, 2016). A Figura 3 ilustra o funcionamento geral de um modelo de aprendizado baseado no método de *boosting*.

Figura 3 – Diagrama de blocos do método *boosting*

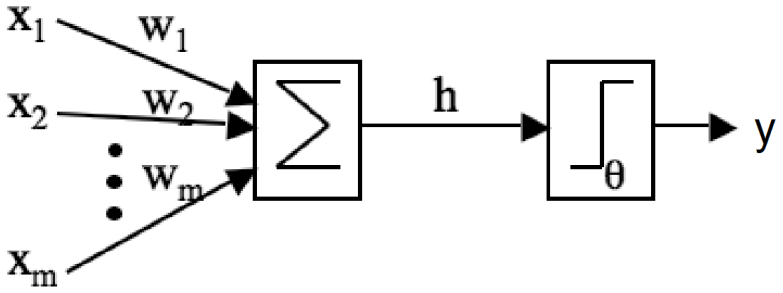


Fonte: Autoria própria

2.4.4 Redes Neurais Artificiais

Para o entendimento do terceiro modelo utilizado neste trabalho (*MLPRegressor*), é necessário compreender os conceitos clássicos que deram origem às redes neurais artificiais (RNAs). Nesse sentido, destaca-se o impacto da obra de McCulloch e Pitts (1943), onde se introduziu pela primeira vez o conceito matemático de um neurônio biológico, cujo esquema é representado na Figura 4.

Figura 4 – Representação do modelo matemático de um neurônio por McCulloch e Pitts



Fonte: Adaptado de Marsland (2015)

O modelo é composto por três seções principais: 1) uma série de entradas x_i, cada uma multiplicada por um peso w_i, representando as sinapses; 2) uma somatória h para todos esses sinais ponderados de entrada, análoga à membrana de uma célula que recebe pulsos elétricos; e 3) uma função de ativação, que utiliza um limiar θ para determinar se o neurônio será ativado, representada pela Equação 2.4. Os valores de entrada provêm de outros neurônios, e o valor de saída y é passado a neurônios subsequentes (MARSLAND, 2015).

$$y = g(h) = \begin{cases} 1 & \text{se } h > \theta \\ 0 & \text{se } h \leq \theta \end{cases} \tag{2.4}$$

2.4.4.1 Perceptron

O Perceptron é a forma mais simples de configuração de uma RNA, constituída por um único neurônio derivado do neurônio de MCCulloch e Pitts, em apenas uma camada. Este recebe um conjunto de valores de entrada x_i, cada um associado a um peso sináptico w_i, onde i representa a entrada associada ao peso. A saída do neurônio y é definida por:

$$\begin{cases} y = g(h) \\ h = \sum_{i=1}^{n} w_i x_i \end{cases} \tag{2.5}$$

Durante o treinamento, o valor de saída y, gerado para a k-ésima amostra de entrada $x^{(k)}$, é comparado à saída desejada $d^{(k)}$. Se houver discrepância, os pesos w_i precisam ser reajustados, já que os valores de entrada x_i e a forma da função $g(h)$ são fixos. A atualização desses pesos é dada por:

$$\Delta w_i = \eta(d^{(k)} - y)x_i^{(k)} \tag{2.6}$$

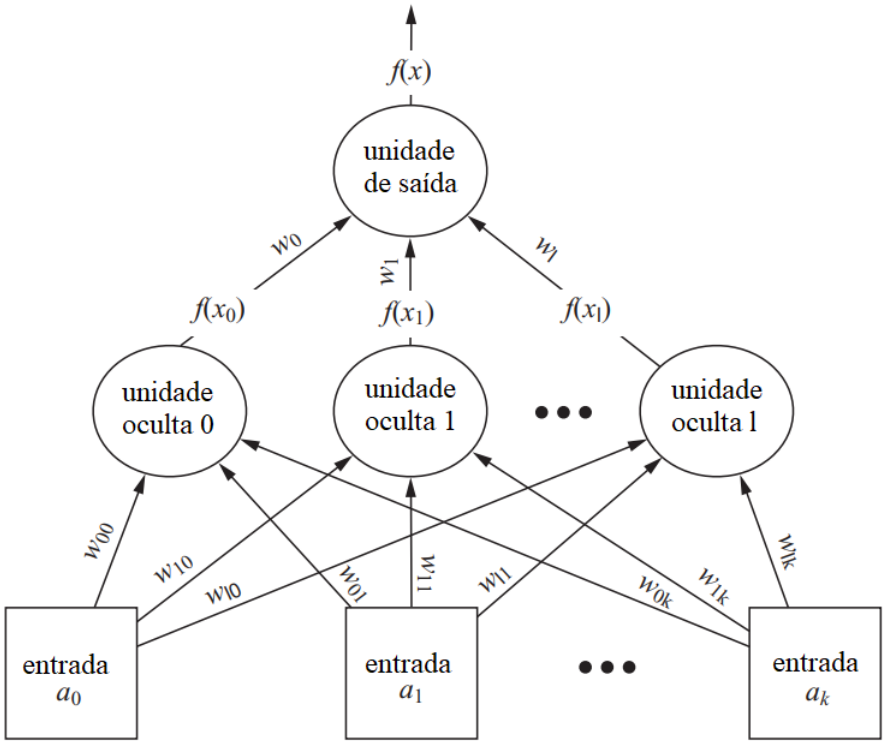
onde Δw_i define o ajuste necessário ao peso w_i e η representa a taxa de aprendizado, que determina o quão rápido a rede irá convergir (i.e., a velocidade de ajuste de w). Marsland (2015) menciona também a inclusão de um termo de viés (*bias*) — uma entrada adicional com valor fixo em 1 — útil para permitir a ativação quando todas as outras entradas são iguais a zero. Seu valor é igualmente atualizado a cada iteração de correção.

2.4.4.2 Perceptron Multicamadas

Witten e Frank (2005) apontam a dificuldade enfrentada por um simples perceptron ao lidar com problemas não linearmente separáveis, pois o algoritmo falha ao tentar encontrar um hiperplano que separe adequadamente os dados. A solução proposta é a introdução dos perceptrons de múltiplas camadas (MLPs), cuja arquitetura propõe a organização dos neurônios em camadas interligadas, permitindo a construção de funções de decisão não lineares. Com isso, os MLPs aumentam significativamente sua capacidade de representação do modelo, alcançando desempenho comparável às DTs, por exemplo.

Sua estrutura é composta por três tipos de camadas: uma camada de entrada, uma ou mais camadas ocultas e uma camada de saída. A camada de entrada recebe os atributos do conjunto de dados; as camadas ocultas são responsáveis por processar e transformar a informação; e a camada de saída fornece o valor estimado. Em cada neurônio oculto, as entradas são multiplicadas por pesos, somadas a um viés e processadas por uma função de ativação. Essa estrutura compõe o fluxo *feedforward*, onde os dados se propagam em um único sentido — da entrada para a saída — sem ciclos ou realimentações entre as camadas. A Figura 5 representa um esquemático de um MLP, onde $f(x_i)$ representa a saída da i-ésima unidade oculta, w_{ij} o peso da conexão da entrada a e a unidade oculta, e w_i o peso da conexão da unidade oculta i para a unidade de saída (WITTEN; FRANK, 2005).

Figura 5 – RNA Perceptron Multicamadas



Fonte: Adaptado de Witten e Frank (2005)

Essa função de ativação aplicada na camada oculta é crucial para a modelagem das relações não lineares entre as variáveis. Dentre as opções mais usadas, destacam-se as funções *sigmoid* e *tanh*, comuns em tarefas de classificação por retornarem valores em intervalos limitados. A função *ReLU* (*Rectified Linear Unit*), por outro lado, ganhou maior notoriedade devido à sua simplicidade computacional e bom desempenho. Em regressões, como o tratado neste trabalho, é comum o uso de uma função de ativação linear (ou nenhuma ativação) na camada de saída, permitindo que o modelo produza valores contínuos sem restrições. Conforme apontado por Aggarwal (2018), essa escolha é essencial para garantir que o modelo possa estimar variáveis numéricas sem distorções introduzidas por funções não lineares.

O aprendizado da rede depende do processo de otimização dos pesos sinápticos, para que as saídas previstas se aproximem das saídas esperadas. Para isso, utiliza-se um algoritmo denominado *gradient descent*, cuja função é minimizar o erro entre as saídas previstas pela rede e os valores reais esperados. A taxa de aprendizado (*learning rate*) controla a magnitude desses ajustes. Para o ajuste afetar todas as camadas do modelo, utiliza-se o algoritmo de *backpropagation*, que consiste em computar o erro a partir da camada de saída e retropropagá-lo até as camadas anteriores, atualizando os pesos com

base nas derivadas parciais da função de ativação. Essa retroalimentação do erro permite que a rede otimize seus parâmetros eficientemente, buscando melhorar progressivamente a acurácia das previsões.

2.4.5 MLPRegressor

O *MLPRegressor* é um algoritmo voltado para a implementação prática das redes neurais do tipo perceptron multicamadas, aplicado especificamente em problemas de regressão. Disponível por fácil acesso na biblioteca *scikit-learn*, o modelo utiliza a mesma arquitetura descrita na seção anterior — com entradas, uma ou mais camadas ocultas e uma camada de saída — para aprender as relações entre variáveis de entrada e valores contínuos de saída. Internamente, a rede é treinada com a retropropagação do erro, empregando métodos de otimização como o *stochastic gradient descent* (SGD) ou *Adam*, que ajustam os pesos sinápticos com base na minimização do erro da função de perda (PEDREGOSA et al., 2011).

Diferentemente das técnicas supracitadas, como o RF ou *XGBoost*, o *MLPRegressor* não é baseado em árvores de decisão e, portanto, consegue capturar padrões não lineares nos dados. Entretanto, por ser uma rede neural, o modelo é extremamente sensível à escolha de hiperparâmetros como o número de camadas ocultas, o número de neurônios por camada, a taxa de aprendizado e a função de ativação. Seu desempenho também pode variar consideravelmente caso os dados de entrada não sejam previamente normalizados, devido à sensibilidade do modelo com a escala das variáveis.

2.5 Métricas de Avaliação para Modelos de Regressão

Este trabalho busca, num estágio inicial, replicar o índice de risco de fogo do INPE. Portanto, para a avaliação dos modelos preditivos desenvolvidos, escolheram-se métricas que quantificam a relação entre os valores previstos pelo modelo e os valores reais medidos. Conforme apontado por Hastie, Tibshirani e Friedman (2009), métricas como o R^2, erro médio absoluto, do inglês *mean absolute error* (MAE), erro quadrático médio, do inglês *mean square error* (MSE), raiz do erro quadrático médio, do inglês *root mean square error* (RMSE) e erro percentual absoluto médio simétrico, do inglês *symmetric mean absolute percentage error* (SMAPE) são amplamente utilizadas e reconhecidas na avaliação de modelos de regressão, e serão adotadas ao longo deste projeto. Seu funcionamento e suas formulações matemáticas estão descritos abaixo, utilizando as seguintes notações: y_i representa o valor observado para a i-ésima amostra, $\hat{y}_i$ corresponde ao valor previsto pelo

modelo para essa mesma amostra, $\bar{y}$ indica a média dos valores observados, e n denota o número total de amostras analisadas.

- **Coeficiente de Determinação (R^2)**
  O R^2 mede a proporção da variabilidade dos dados que é explicada pelo modelo. A métrica compara a soma dos erros quadráticos do modelo com a variância total do valor real. O resultado varia de $-\infty$ a 1, sendo que valores mais próximos de 1 indicam um melhor ajuste. Um valor $R^2 = 0$ indica que o modelo não apresenta nenhuma variância dos dados, enquanto valores negativos significam que o modelo é pior do que um modelo que sempre prediz a média dos valores reais. É dado por:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{2.7}$$

- **Erro absoluto médio (MAE)**
  O MAE corresponde à média dos valores absolutos das diferenças entre as previsões e os valores reais. É uma métrica direta e interpretável, pois se encontra na mesma unidade dos dados analisados. A MAE é especialmente útil em contextos onde não se deseja penalizar desproporcionalmente erros maiores, ao contrário do que ocorre com o erro quadrático médio (MSE). É dado por:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{2.8}$$

- **Erro quadrático médio (MSE)**
  O MSE representa a média dos quadrados das diferenças entre os valores reais e previstos. Devido à elevação ao quadrado, penaliza fortemente desvios maiores, sendo sensível a *outliers*. Essa característica torna o MSE útil quando erros maiores são especialmente indesejados. A unidade da métrica é o quadrado da unidade dos dados, dificultando a interpretação direta. É dado por:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{2.9}$$

- **Raiz do erro quadrático médio (RMSE)**
  A RMSE é a raiz quadrada do MSE, retornando os erros à mesma unidade dos valores de entrada. Essa métrica também penaliza fortemente grandes erros, mas sua escala torna os resultados mais intuitivos. É frequentemente usada em aplicações práticas por equilibrar interpretabilidade com sensibilidade a erros extremos. É dada por:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{2.10}$$

- **Erro percentual absoluto médio simétrico (SMAPE)**
  O SMAPE é uma métrica percentual baseada na razão entre o erro absoluto e a média dos valores absolutos real e previsto. É útil por padronizar o erro, permitindo comparação entre séries de diferentes escalas. Seu valor varia de 0 a 100%, sendo 0 o erro ideal. A simetria do denominador evita penalizações desbalanceadas quando os valores reais ou previstos são muito próximos de zero. É dado por:

$$\text{SMAPE} = \frac{100\%}{n} \sum_{i=1}^{n} \frac{|y_i - \hat{y}_i|}{(|y_i| + |\hat{y}_i|)/2} \tag{2.11}$$

As métricas apresentadas serão aplicadas na avaliação comparativa dos modelos, servindo de base para a análise de desempenho e seleção da abordagem mais eficaz para o problema proposto.

3 Revisão Bibliográfica

O desenvolvimento de índices de risco de fogo, com base em dados topográficos, climatológicos e em imagens de alta resolução geradas por satélite, não se limita ao Programa Queimadas do INPE (2025). Em outros países, iniciativas similares também já foram adotadas, como o Drought and Fire Observatory and Early Warning System (2025), European Forest Fire Information System (2025), Fire Information for Resource Management System US / Canada (2025), Ontario: Forest Fire Info Map (2025) e Canadian Forest Fire Weather Index (2025).

Dentre as metodologias aplicadas, Setzer, Sismanoglu e Santos (2019) descrevem o método utilizado para cálculo do risco de fogo dentro do Programa Queimadas (que constitui o objeto de estudo deste trabalho). Este baseia-se principalmente nos "Dias de Secura" (PSE), referentes ao número de dias seguidos sem nenhuma precipitação durante os últimos 120 dias. Outros fatores, como o tipo de vegetação, temperatura máxima, umidade relativa mínima do ar, elevação topográfica e presença do fogo também são considerados no cálculo do RF, que indica o grau de propensão da vegetação em ser queimada do ponto de vista meteorológico. Similarmente, o cálculo do FWI (*Canadian Forest Fire Weather Index*, índice canadense de risco meteorológico de incêndios florestais) leva em conta observações diárias de temperatura, umidade relativa, vento e precipitação para definição de três *fuel moisture codes* — índices que servem de base para o cálculo do FWI e representam o teor de umidade do solo florestal e de outras matérias orgânicas mortas (CWFIS, 2025).

Não é incomum, na literatura, a aplicação de ML em cenários de previsão de eventos associados a fenômenos meteorológicos ou condições climáticas. Um dos trabalhos pioneiros relacionados à obtenção da correlação entre a análise de dados meteorológicos e a área queimada é o de Cortez e Morais (2007). Utilizando somente quatro variáveis de entrada — chuva, temperatura, umidade e velocidade do vento — os autores desenvolveram um modelo capaz de prever a extensão das áreas queimadas no Parque Nacional de Montesinho, em Portugal. O estudo previu aproximadamente 46% dos focos com erro inferior a $1ha$, e esse número subia para 61% quando o erro admitido era de até $2ha$. A pesquisa foi uma das primeiras iniciativas bem-sucedidas em estimar a área queimada de uma região com base exclusivamente em variáveis meteorológicas.

Coutinho, Silva e Delgado (2016) propõem a utilização de RNAs de arquitetura *feedforward*, juntamente com algoritmos de treinamento de *backpropagation* e de funções de base radial (RBFs) para predição de dados meteorológicos na região de Paty do Alferes

e Paracambi, no Rio de Janeiro, comparados posteriormente com a aplicação de modelos de regressão linear múltipla (RLM). Os resultados obtidos apresentaram entre 91% e 96% de acerto em todos os casos.

O estudo em Rubí e Gondim (2024) analisa a aplicação de diversos modelos de ML para predição do índice de queimadas na região do Distrito Federal. Dentre os oito modelos testados (RNA, máquina de vetores de suporte, do inglês *support vector machine* (SVM), RF, classificador ingênuo de Bayes com distribuição gaussiana (*Gaussian Naive Bayes*), KNN, regressão linear, regressão logística (LogR) e *AdaBoost*), RF e *AdaBoost* apresentaram os melhores resultados. Este, por sua vez, registrou a melhor média de área sob a curva característica de operação do receptor (ROC, do inglês *receiver operating characteristic*) com 0,993.

Similarmente, Gholamnia et al. (2020) realizou a comparação entre 11 métodos diferentes de ML, aplicando-os para avaliação da suscetibilidade de queimadas no condado de Amol, na província de Mazandaran, Irã. Foram escolhidos os métodos de RNA, regressão *Dmine* (DR, do inglês *Dmine regression*), mineração de dados neurais, LARS (regressão por ângulo mínimo, do inglês *least angle regression*), MLP, RF, RBF, Mapa de Kohonen, SVM, DT e LogR. Categorizaram-se 66% dos dados para treinamento e 33% para validação, com três divisões de validação cruzada. A curva ROC foi utilizada para avaliar a precisão das abordagens. Dentre todos, RF demonstrou-se como a melhor opção para predição de incêndios florestais, apresentando uma precisão de 88%.

Em outro estudo aplicado sobre biomas brasileiros, (GALIZIA; RODRIGUES, 2019) procurou prever e detectar mudanças nos padrões de ignição de queimadas no Cerrado brasileiro, em decorrência das plantações de eucalipto. Aplicou, para tal, a combinação de algoritmos de RF com análise em *clusters*. A variável dependente foi a ausência ou não de fogo na região. O modelo apresentou precisão de 0,75 em área sob curva ROC.

Oliveira et al. (2022) propõem um modelo para análise do impacto do fogo nos biomas brasileiros, produzindo um índice baseado em cinco características do fogo (intensidade, recorrência, extensão, intervalo entre fogos e sazonalidade). Para validar a abordagem, os autores o compararam com diversos modelos, tal como o SAR, Generalized Linear Model (GLM), SVM e RF. Os resultados indicaram que o modelo baseado em RF apresentou o melhor desempenho na estimativa do impacto do fogo, com resultados em média 70% melhores que o segundo colocado.

A partir da análise dos trabalhos revisados, constata-se a ampla aplicação de técnicas de ML na predição de incêndios florestais, com destaque para os modelos de RF, reconhecidos por sua robustez e boa capacidade de generalização. Diante disso, o presente trabalho propõe a construção e avaliação de três modelos preditivos — RF, *MLPRegressor*

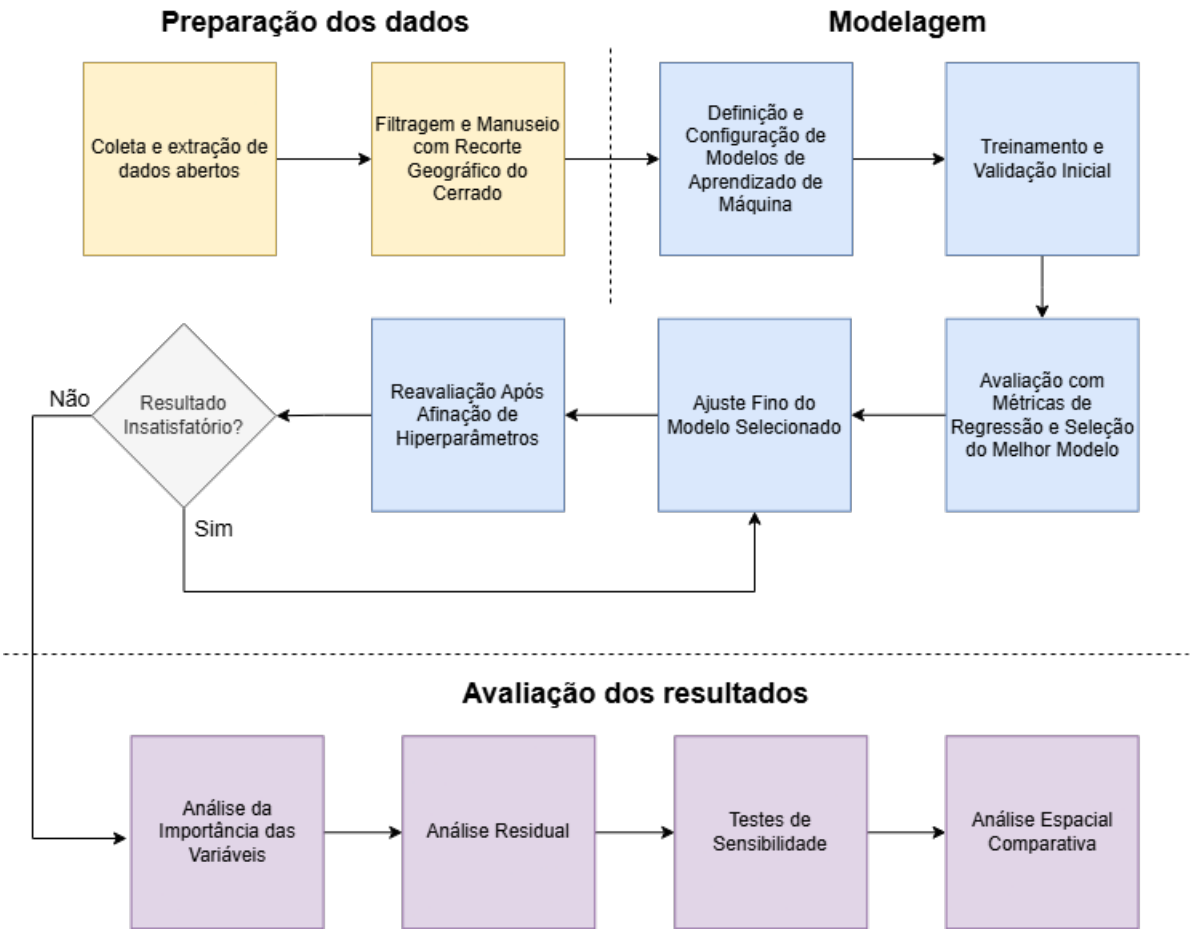
e *XGBoost* — baseados em um conjunto reduzido de variáveis meteorológicas e atmosféricas, com foco na replicação do índice de risco de fogo do INPE para a região do Cerrado.

Embora os estudos revisados concentrem-se predominantemente no desempenho dos modelos preditivos, este trabalho busca também incluir a análise da contribuição individual de cada variável nas predições geradas, oferecendo subsídios adicionais para a interpretação do fenômeno e a construção de políticas de prevenção. A partir dos resultados obtidos, o modelo com melhor desempenho será utilizado para análise da importância das variáveis e para avaliação da eficácia de diferentes combinações preditoras, estabelecendo uma base para estudos posteriores e tomadas de decisão. Os trabalhos revisados serão retomados no capítulo 5 para fins comparativos, principalmente em relação a relevância atribuída às variáveis de entrada.

4 Desenvolvimento do Trabalho

Este capítulo apresenta as etapas práticas adotadas na construção dos modelos preditivos propostos. Para facilitar a compreensão do fluxo metodológico adotado, elaborou-se um diagrama de blocos representado na Figura 6. O diagrama está dividido em três seções principais: preparação dos dados, modelagem e avaliação dos resultados. Cada bloco do diagrama representa uma etapa específica, conectada conforme a sequência lógica adotada no projeto. As seções a seguir explicam detalhadamente as atividades desenvolvidas em cada uma dessas etapas.

Figura 6 – Diagrama de blocos da metodologia aplicada



Fonte: Autoria própria

4.1 Preparação dos Dados

A primeira seção consistiu na preparação e no tratamento dos dados. Essa fase envolveu a identificação e coleta de variáveis ambientais disponibilizadas publicamente pela seção de Dados Abertos do Programa Queimadas do INPE, seguida da filtragem de amostras válidas, da padronização para uma mesma resolução espacial e da delimitação geográfica correspondente ao bioma Cerrado. As subseções a seguir detalham cada uma dessas etapas.

4.1.1 Coleta e Extração de Dados Abertos

A partir da base de dados definida, conforme mencionado anteriormente no Capítulo 2, selecionaram-se dados com resolução diária referentes à temperatura média, umidade relativa do ar, precipitação acumulada, número de dias consecutivos sem chuva e concentração atmosférica de fumaça. A tabela 1 resume as informações referentes às variáveis independentes (*features*) e à variável dependente (*target*) consideradas neste trabalho.

Tabela 1 – Conteúdo das variáveis do conjunto de dados

Descrição	**Tipo de variável**	**Unidade de medida**	**Formato**
Temperatura do ar	*feature*	K	*.nc*
Umidade relativa do ar	*feature*	%	*.nc*
Precipitação acumulada	*feature*	mm/dia	*.nc*
Número de dias consecutivos sem chuva	*feature*	dias	*.nc*
Concentração atmosférica de fumaça	*feature*	μg/m^3	*.tiff*
Índice do risco de fogo	*target*	adim. (0 a 1)	*.nc*

Cada variável é disponibilizada pelo portal com frequência diária. Para o tratamento, organização e modelagem dos dados, empregaram-se ferramentas amplamente reconhecidas pela comunidade científica pela flexibilidade, eficiência com grandes volumes de dados e disponibilidade gratuita. O ambiente de desenvolvimento adotado foi o *Anaconda* (Anaconda Inc., 2025), com o uso do editor *Spyder* (Spyder IDE, 2025) e a linguagem de programação *Python* (Python Software Foundation, 2025).

A biblioteca *netCDF4* foi empregada para manuseio dos dados diários de formato *.nc* (*NetCDF — Network Common Data Form*), estruturando-os como um *Dataset* tridimensional: a primeira dimensão corresponde ao tempo (um único instante por arquivo); a segunda à latitude; e a terceira à longitude. Apesar dos arquivos compartilharem a mesma estrutura, as dimensões espaciais (latitude x longitude) variam entre as variáveis.

Algumas variáveis, como o índice de risco de fogo, apresentam maior resolução espacial em comparação às demais.

A única variável disponibilizada em outro formato foi a concentração atmosférica de fumaça — organizada em arquivos *.tiff* (*GeoTIFFs*). Para o manuseio da mesma, foi utilizada a biblioteca *rasterio*, que permitiu o mapeamento das imagens em coordenadas geográficas reais. A tabela 2 apresenta as dimensões espaciais de cada variável, bem como os limites de latitude e longitude observados.

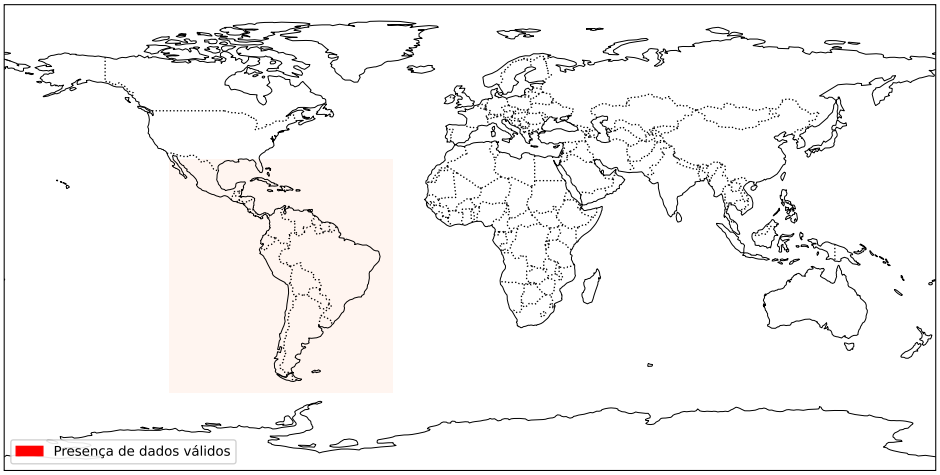
Tabela 2 – Dimensão das variáveis utilizadas no conjunto de dados

Variável	**Resolução (lat x lon)**	**Latitude (extremos)**	**Longitude (extremos)**
Temperatura do ar	(361, 345)	[-60,30]	[-116, -30]
Umidade relativa do ar	(361, 345)	[-60,30]	[-116, -30]
Precipitação acumulada	(901, 850)	[-60.05, 29.95]	[-114.95, -30.05]
Número de dias consecutivos sem chuva	(901, 850)	[-60.05, 29.95]	[-114.95, -30.05]
Concentração atmosférica de fumaça	(231, 231)	[-59, 33]	[-120, -28]
Índice do risco de fogo	(8899, 8899)	[-55.985, 32.995]	[-119.99, -33.0098]

4.1.2 Filtragem e Manuseio com Recorte Geográfico do Cerrado

Os dados disponibilizados pelo Programa Queimadas extrapolam os limites territoriais do Brasil, abrangendo praticamente toda a América Latina, conforme observado na Figura 7. Visando otimizar o processamento computacional e evitar o treinamento desnecessário de modelos sobre regiões irrelevantes para este estudo, implementou-se um filtro espacial preliminar ao processamento dos dados através da definição de um "recorte retangular", que engloba toda a extensão do bioma Cerrado, com base em intervalos mínimos e máximos de latitude e longitude. Embora o recorte não represente precisamente a delimitação oficial do bioma, ele permite uma redução significativa da área processada, diminuindo em aproximadamente 94% o volume dos dados analisados e acelerando a execução dos códigos sem comprometer a cobertura da região de interesse.

Figura 7 – Cobertura espacial dos dados ambientais disponíveis

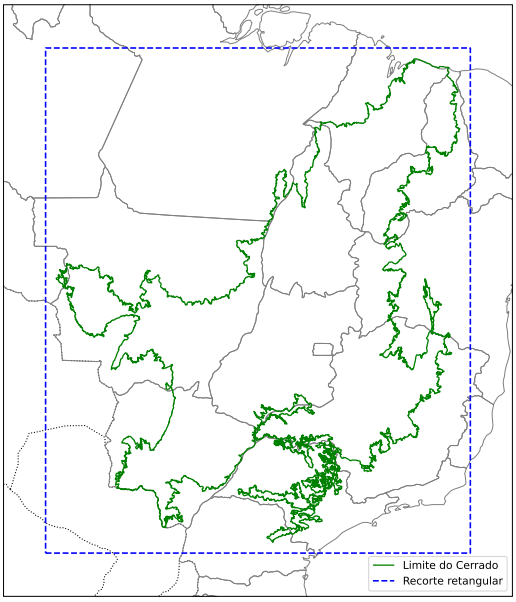


Fonte: Autoria própria

Paralelamente, definiram-se também os limites espaciais próximos à extensão real do Cerrado. Para isso, utilizou-se a biblioteca *geopandas*, que possibilita operações espaciais baseadas em dados vetoriais. O portal INPE (2025) disponibiliza gratuitamente arquivos no formato *.shp* (*Shapefile*) contendo os contornos georreferenciados dos biomas brasileiros. A partir do arquivo correspondente ao Cerrado, foi realizada uma interseção entre os pontos de dados de um *Dataset* de referência e o polígono "oficial" do bioma.

A escolha dessa variável de referência baseou-se em dois critérios: (i) estar no mesmo formato da maioria dos arquivos utilizados (no caso, *.nc*); e (ii) possuir a menor resolução espacial entre os disponíveis. Este segundo critério foi adotado para minimizar a necessidade de *upscaling*, isto é, da interpolação de dados para uma resolução superior, o que poderia acarretar a introdução de ruídos ou distorções. Com base nesses critérios, a variável **temperatura** foi selecionada como referencial espacial para o tratamento dos dados. A Figura 8 ilustra a diferença entre o recorte retangular inicial e a delimitação precisa do Cerrado, em contraste com a extensão do território brasileiro.

Figura 8 – Delimitação geográfica do bioma Cerrado e área de análise



Fonte: Autoria própria

Subsequente ao filtramento inicial baseado no "recorte retangular", procedeu-se à remoção de amostras inválidas, descartando valores ausentes (*NaN* ou marcados com -999, associados a falhas na coleta dos dados). Em seguida, realizou-se a correspondência entre as coordenadas (latitude e longitude) da variável de referência e das demais variáveis analisadas. Para os casos em que os valores não estavam diretamente disponíveis, aplicaram-se técnicas de interpolação espacial, conforme discutido anteriormente no Capítulo 2.

A biblioteca *SciPy*, por meio da função *griddata*, disponibiliza métodos de interpolação baseados em triangulação de Delaunay. Conforme a documentação, essa abordagem é especialmente indicada para malhas espaciais com distribuição irregular de pontos, apresentando boa acurácia em variáveis contínuas. Essa característica a torna particularmente adequada ao contexto deste estudo, onde os dados ambientais nem sempre estão uniformemente distribuídos (SCIPY, 2023).

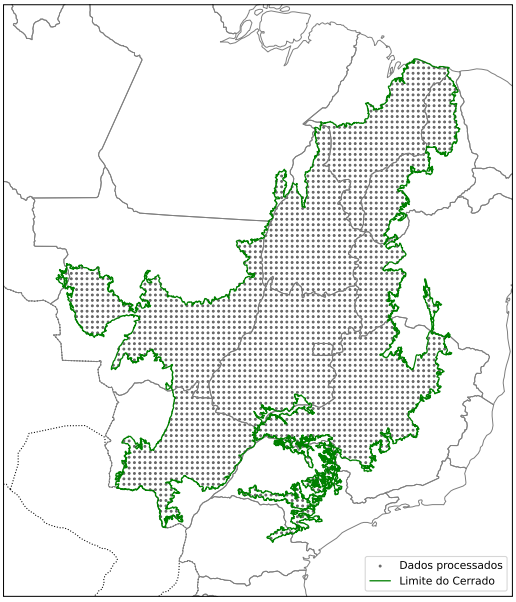
Durante o desenvolvimento do trabalho, contudo, observou-se que a aplicação da interpolação via *griddata* em todos os conjuntos de dados era inviável, principalmente devido ao elevado tempo de processamento computacional, sobretudo no caso de variáveis com alta resolução espacial. Em especial, para o índice de risco de fogo — que, conforme a Tabela 2, apresenta resolução significativamente superior às demais variáveis —, optou-se por uma alternativa escalável.

Nesses casos, portanto, a interpolação foi realizada utilizando o algoritmo KNN (k vizinhos mais próximos) com ponderação pelo inverso da distância (IDW), sendo os vizinhos determinados por meio da estrutura *cKDTree*, disponibilizada também pela biblioteca *SciPy*. Essa estrutura fornece um índice para voltado à busca em um conjunto de pontos k-dimensionais, que permite localizar rapidamente os vizinhos mais próximos de qualquer ponto (SCIPY, 2023).

Todos os processos descritos anteriormente são aplicáveis tanto para arquivos de formato *.nc* quanto para *.tiff*, com variações mínimas no manuseio e na leitura dos dados. Essas diferenças foram facilmente tratadas graças à diversidade de bibliotecas disponíveis na linguagem *Python*. Os códigos desenvolvidos neste projeto foram estruturados de forma modular, priorizando a segmentação das execuções para permitir maior controle sobre os resultados gerados em cada etapa.

A última etapa desta fase consistiu no processamento diário das variáveis, seguido do armazenamento dos dados filtrados em arquivos no formato *Parquet*. Esse formato foi escolhido, principalmente, por sua eficiência na leitura e escrita de grandes volumes de dados tabulares. Cada dia de coleta originou um arquivo *Parquet* correspondente, contendo todas as informações espacialmente válidas no Cerrado. É importante destacar que, caso os dados de qualquer uma das variáveis não estivessem disponíveis em um determinado dia, esse dia era descartado e não integrado ao conjunto final dos dados. A Figura 9 ilustra os limites geográficos do Cerrado e os pontos considerados válidos a partir da variável de referência, após a filtragem espacial.

Figura 9 – Dados processados nos limites do Cerrado



Fonte: Autoria própria

Ao final deste processo, foram gerados 653 arquivos diários, contendo valores válidos para todas as seis variáveis consideradas, restritos exclusivamente à região do Cerrado. Originalmente, a variável de referência utilizada (temperatura) apresenta uma resolução espacial de [361 × 345] pontos — totalizando 124.545 combinações de latitude e longitude em cada arquivo diário. Após o recorte geográfico, restaram 2.652 pontos válidos, ou seja, localizações efetivamente situadas nos limites do Cerrado com dados disponíveis para todas as variáveis, representando cerca de 2% do total original. A unificação desses arquivos diários ao longo do intervalo temporal (de agosto de 2023 até maio de 2025) constitui a base final utilizada na etapa de modelagem.

4.2 Modelagem

Com os dados preparados, iniciou-se a modelagem preditiva por meio da aplicação de técnicas de aprendizado de máquina. Nessa etapa, foram definidos os modelos a serem avaliados, realizadas as fases de treinamento e validação, e selecionado aquele com melhor desempenho preditivo com base em métricas específicas para problemas de regressão. A partir dessa seleção, procedeu-se ao ajuste fino dos hiperparâmetros visando otimizar os resultados, até que os mesmos fossem satisfatórios. As etapas e critérios adotados são apresentados a seguir.

4.2.1 Definição e Configuração de Modelos de Aprendizado de Máquina

Conforme discutido previamente no Capítulo 2, este trabalho propõe a utilização de três modelos distintos de aprendizado de máquina (ML) para a predição do índice do risco de fogo. A seleção foi fundamentada na natureza do problema — regressão — e respaldada tanto pela recorrência desses modelos na literatura quanto por seus desempenhos satisfatórios observados durante a revisão bibliográfica. Os modelos adotados foram: *Random Forest* (RF), *XGBoost* e *MLPRegressor*.

O RF é baseado na técnica de *bagging*, e é amplamente reconhecido por sua robustez e desempenho em problemas compostos por conjuntos de dados extensos e variáveis interdependentes. O *XGBoost*, por sua vez, utiliza o princípio de *boosting*, ajustando iterativamente novos modelos para corrigir os erros residuais das previsões anteriores. Por fim, o *MLPRegressor* representa uma abordagem baseada em redes neurais do tipo MLP, sendo o único modelo entre os selecionados que não se baseia em DT. Sua estrutura permite capturar relações não lineares complexas entre as variáveis de entrada e a variável-alvo.

4.2.2 Treinamento e Validação Inicial

Inicialmente, essa etapa inicial é dedicada à definição de um modelo dentre os três selecionados. O ajuste fino será feito posteriormente, a partir desse modelo único. Por isso, adotou-se uma abordagem clássica e comum para o treinamento e validação, que pudesse ser aplicada de forma genérica em cima dos modelos sugeridos. Baseado no conjunto inicial obtido após o tratamento dos dados válidos, foi feita a separação das amostras, dedicando 80% dos dados para treinamento e 20% para testes e avaliação de métricas.

Para os três algoritmos, utilizou-se uma estratégia de validação cruzada combinada com ajuste de hiperparâmetros. Essa abordagem procura avaliar o desempenho médio dos modelos em subconjuntos distintos dos dados, ao mesmo tempo em que identifica, para cada um deles, a combinação de hiperparâmetros que proporciona os melhores resultados preditivos.

Para o processo de validação cruzada (*cross-validation*), adotou-se especificamente a técnica de *k-fold cross-validation*, com $k = 5$. Esse método consiste em dividir o conjunto de dados em cinco subconjuntos de tamanhos semelhantes. A cada iteração, quatro desses subconjuntos são utilizados para o treinamento do modelo, enquanto o subconjunto restante é reservado para teste. Repete-se o procedimento cinco vezes, alternando o subconjunto de teste. O desempenho final do modelo é obtido por meio da média das métricas observadas nas cinco iterações. Essa técnica é amplamente reconhecida por fornecer estimativas mais robustas de desempenho do modelo, especialmente em casos em que os dados não são homogêneos (HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

A implementação dessa abordagem, aliada ao ajuste de hiperparâmetros, foi realizada por meio da ferramenta *GridSearchCV*, disponibilizada pela biblioteca *scikit-learn*. A ferramenta realiza uma busca exaustiva entre as combinações de hiperparâmetros pré-definidos, avaliando cada configuração a partir da validação cruzada aplicada sobre os dados de treinamento. Dessa forma, é possível identificar sistematicamente qual conjunto de parâmetros maximiza o desempenho preditivo de cada modelo, considerando tanto a adequação ao problema quanto as condições ideais de operação.

Para o ajuste de hiperparâmetros, cada modelo foi configurado com um conjunto restrito e representativo de combinações, selecionadas para permitir variações significativas sem comprometer o custo computacional. Considerando que esta etapa ainda representa uma triagem inicial para seleção de um modelo mais promissor, definiu-se a avaliação de 4 a 5 hiperparâmetros por algoritmo, cada um com 2 a 3 valores possíveis. A seguir, apresentam-se as combinações de hiperparâmetros testadas para cada modelo, com os valores padrão destacados em negrito, acompanhadas de uma breve explicação sobre cada parâmetro:

RandomForest

- Total de $2 \times 2 \times 2 \times 2 \times 2 = 32$ combinações ($32 \times$ 5-fold $= 160$ modelos treinados)
  - *n_estimators* [**100**, 200]: número de árvores no conjunto da floresta; mais árvores tendem a reduzir a variância.
  - *max_depth* [10, **None**]: profundida máxima de cada árvore; *None* permite o crescimento até as folhas.
  - *min_samples_split* [**2**, 5]: número mínimo de amostras exigido para dividir um nó interno.
  - *min_samples_leaf* [**1**, 2]: número mínimo de amostras exigido para um nó ser considerado folha.
  - *max_features* [**1**, 'sqrt']: número de variáveis consideradas aleatoriamente em cada divisão de uma árvore.

XGBoost

- Total de $3 \times 3 \times 3 \times 2 \times 2 = 108$ combinações ($108 \times$ 5-fold $= 540$ modelos treinados)
  - *n_estimators* [**100**, 200, 300]: número de árvores adicionadas sequencialmente no conjunto da floresta.
  - *max_depth* [5, **6**, 7]: profundida máxima da cada árvore.

  - *learning_rate* [0.01, 0.1, **0.3**]: taxa de aprendizado aplicada à cada nova árvore, reduzindo o peso de cada uma no modelo final.
  - *subsample* [0.8, **1.0**]: taxa de amostras utilizadas em cada iteração de treinamento.
  - *colsample_bytree* [0.8, **1.0**]: proporção de atributos selecionados na construção de cada árvore.

MLPRegressor

- Total de $4 \times 2 \times 3 \times 2 = 48$ combinações ($48 \times$ 5-fold $= 240$ modelos treinados)
  - *hidden_layer_sizes* [(50,), **(100,)**, (50, 50), (100, 50)]: arquitetura da rede, definida pelo número de neurônios em cada camada oculta.
  - *activation* [**'relu'**, 'tanh']: função de ativação aplicada na camada oculta.
  - *alpha* [**0.0001**, 0.001, 0.01]: métrica de regularização L2, controla a complexidade do modelo.
  - *learning_rate_init* [**0.001**, 0.01]: valor inicial da taxa de aprendizado, atualiza os pesos da rede durante o treinamento via *backpropagation*.

Antes da aplicação dos algoritmos, é preciso considerar as particularidades de cada modelo em relação à escalabilidade dos dados. Modelos baseados em árvores de decisão, como o *Random Forest* e o *XGBoost*, não são afetados pela escala dos atributos, pois suas decisões são tomadas baseadas em divisões sucessivas dos dados, considerando apenas a ordem ou os valores relativos das variáveis. Isso torna a escala irrelevante em seu funcionamento (MARSLAND, 2015).

Por outro lado, modelos baseados em RNAs, como o *MLPRegressor*, são sensíveis à escala dos dados de entrada por utilizarem operações com pesos e funções de ativação, cujo comportamento é estável e eficiente quando os dados estão normalizados em uma mesma escala. Para este modelo, aplicou-se a padronização dos dados utilizando o método *StandardScaler()*, da biblioteca *scikit-learn*. A técnica transforma os dados para que possuam média zero e desvio padrão igual a um, acelerando a convergência do algoritmo de otimização e melhorando a estabilidade do treinamento (PEDREGOSA et al., 2011).

É importante destacar também que a diferença no número de modelos treinados para cada algoritmo (160 para RF, 540 para *XGBoost* e 240 para *MLPRegressor*) decorre de observações empíricas sobre o tempo de execução e pela viabilidade computacional notada durante os testes preliminares. O *XGBoost*, por exemplo, mesmo com o maior número de combinações avaliadas, apresentou um tempo de execução para a busca exaustiva de 56 minutos, enquanto o RF e o *MLPRegressor* apresentaram tempos mais elevados mesmo

com menos combinações (3h48 e 2h28, respectivamente), o que motivou uma redução no espaço de busca para esses modelos. Ainda assim, mesmo nos casos com menor número de combinações, a quantidade de modelos treinados foi suficiente para suportar análises comparativas robustas entre os métodos.

4.2.3 Avaliação com Métricas de Regressão e Seleção do Melhor Modelo

A seleção do modelo mais promissor entre os algoritmos treinados envolveu duas etapas principais. Primeiramente, utilizou-se o método *GridSearchCV* para encontrar, em cada um dos três algoritmos avaliados, a melhor combinação de hiperparâmetros com base na métrica de R^2, adotada como critério principal nesta etapa inicial. Essa métrica foi escolhida por seu amplo uso e aceitação na literatura em problemas aplicados à regressão, servindo como um indicador global de desempenho preditivo.

Em seguida, os três modelos com melhores desempenhos em R^2 — um para cada algoritmo — foram comparados utilizando as demais métricas discutidas no Capítulo 2: MAE, MSE, RMSE e SMAPE. O modelo com o melhor desempenho agregado nessas métricas foi então selecionado para posterior ajuste fino. A seguir, apresentam-se os três modelos selecionados, juntamente com suas respectivas configurações de hiperparâmetros definidas por *GridSearchCV*.

RandomForest

- *n_estimators*: 200
- *max_depth*: None
- *min_samples_split*: 5
- *min_samples_leaf*: 2
- *max_features*: 'sqrt'

XGBoost

- *n_estimators*: 300
- *max_depth*: 7
- *learning_rate*: 0.1
- *subsample*: 0.8
- *colsample_bytree*: 1.0

MLPRegressor

- *hidden_layer_sizes*: (100, 50)
- *activation*: 'relu'
- *alpha*: 0.0001
- *learning_rate_init*: 0.001

A Tabela 3 apresenta o desempenho de cada um dos três algoritmos com os ajustes de hiperparâmetros supracitados. Os melhores resultados para cada métrica estão destacados em negrito. As métricas de regressão fornecem a base comparativa entre as abordagens e, com base nelas, optou-se pela adoção do modelo *RandomForest*, por apresentar superioridade em todas as métricas avaliadas.

Tabela 3 – Avaliação dos modelos no conjunto de teste por métricas de regressão

Modelo	R^2	**MAE**	**MSE**	**RMSE**	**SMAPE**
RandomForest	**0,8152**	**0,1107**	**0,0352**	**0,1876**	**84,39%**
XGBoost	0,8121	0,1130	0,0358	0,1891	88,35%
MLPRegressor	0,8082	0,1145	0,0365	0,1911	88,68%

4.2.4 Ajuste Fino do Modelo Selecionado

Para o ajuste fino do modelo, optou-se inicialmente por uma segunda rodada de *GridSearchCV*, com redução do espaço de busca e foco nas regiões vizinhas da combinação de hiperparâmetros obtida na etapa anterior. Com base nos resultados desta nova otimização, ajustes manuais e pontuais foram realizados em parâmetros específicos.

Os novos valores foram definidos com a intenção de explorar pequenas variações em torno da configuração considerada ótima, por meio de uma estratégia conservadora voltada ao refino ao invés da expansão do espaço de busca. Adicionalmente, a abordagem evita reintroduzir combinações já descartadas, reduzindo o custo computacional por consequência.

Especificamente, o número de estimadores (*n_estimators*) foi mantido próximo de 200, testando valores como 150 e 250. Para a profundidade máxima das árvores (*max_depth*), ampliou-se o intervalo de valores para avaliação do impacto dos limites. Os parâmetros *min_samples_split* e *min_samples_leaf*, definidos anteriormente como 5 e 2, respectivamente, também foram ajustados para intervalos de valores vizinhos, testando a sensibilidade do modelo. Por fim, quanto ao número de atributos considerados em cada divisão (*max_features*), evitou-se o valor padrão igual a 1, pois este considera apenas uma

variável por divisão. Como o problema depende de múltiplas entradas, adicionou-se a opção *'log2'*, parâmetro apropriado para problemas com a presença de entradas correlacionadas e potencialmente complementares. Em resumo, as combinações possíveis nessa segunda rodada foram testadas com foco em ajustes finos, com os valores do modelo otimizado destacados em negrito e apresentados abaixo:

RandomForest

- Total de $3 \times 2 \times 3 \times 3 \times 2 = 108$ combinações ($108 \times$ 5-fold $= 540$ modelos treinados)
  - *n_estimators*: [150, **200**, 250]
  - *max_depth*: [**None**, 20]
  - *min_samples_split*: [4, **5**, 6]
  - *min_samples_leaf*: [1, **2**, 3]
  - *max_features*: [**'sqrt'**, 'log2']

4.2.5 Reavaliação após Reajuste de Hiperparâmetros

Como resultado da etapa anterior, o modelo que obteve o melhor desempenho em termos da métrica de R^2, tendo sido selecionado durante o processo de busca por meio de *GridSearchCV*, apresentou a seguinte combinação otimizada de hiperparâmetros:

RandomForest

- *n_estimators*: 250
- *max_depth*: 20
- *min_samples_split*: 4
- *min_samples_leaf*: 2
- *max_features*: 'log2'

Algumas observações relevantes podem ser feitas a partir dessa segunda rodada de validação. Em primeiro lugar, os modelos continuam apresentando melhor desempenho com o valor de *min_samples_leaf* igual a 2, demonstrando a estabilidade desse parâmetro. A substituição de *max_features='sqrt'* por *'log2'* também se mostrou eficaz, indicando que outros critérios de seleção de atributos também devem ser considerados. No caso de *min_samples_split*, os valores testados apontam para uma faixa de desempenho ideal observada

entre os valores 2 e 5 (testados na primeira rodada). Para *n_estimators*, observou-se que o aumento do número de árvores — de 200 para 250 — resultou em leve melhora no desempenho. Essa preferência por valores maiores observada rodada a rodada é esperada, dado que um número maior de árvores contribui para a redução da variância do modelo. Por outro lado, a seleção de *max_depth*=20 foi inesperada, considerando que o valor *None* (isto é, sem limitação de profundidade) havia sido mais eficaz anteriormente. Esse comportamento será investigado em uma rodada adicional.

A Tabela 4 apresenta os desempenhos obtidos pelos modelos da primeira e segunda rodada de otimização de hiperparâmetros, com os melhores resultados destacados em negrito. A última linha mostra a variação percentual entre os dois modelos.

Tabela 4 – Desempenho comparativo dos modelos após primeira e segunda rodada de otimização

Modelo	R^2	**MAE**	**MSE**	**RMSE**	**SMAPE**
1º	0,8152	**0,1107**	0,0352	0,1876	**84,39%**
2º	**0,8165**	0,1108	**0,0349**	**0,1869**	87,65%
Δ%	0,15%	0,09%	0,85%	0,37%	3,86%

Entre o primeiro e o segundo modelo, observa-se pouca melhora no desempenho preditivo, refletida pela baixa diferença relativa em todas as métricas de regressão avaliadas. A maioria delas apresentou variações menores que 1%, enquanto houve regressão significativa na avaliação do SMAPE. Além disso, o tempo necessário para o treinamento da segunda rodada de validação foi significativamente maior, totalizando mais de cinco horas. Diante dessa disparidade entre o custo computacional e o ganho obtido, optou-se por realizar um último ajuste fino, restringindo-se aos parâmetros que ainda não demonstraram convergência para valores otimizados.

4.2.5.1 Rodada Final de Validação

Nessa rodada final, manteve-se o uso do *GridSearchCV*, agora aplicado a um escopo de busca reduzido, com foco em três hiperparâmetros: *max_depth*, que apresentou comportamento divergente em relação à primeira rodada; *min_samples_split*, cujo valor ótimo ainda apresenta variação num intervalo estreito (2 a 5); e *n_estimators*, no qual se espera ganho de desempenho com o aumento no número de árvores. *max_features* e *min_samples_leaf* demonstraram estabilidade nas rodadas anteriores, nos intervalos testados, e por isso foram mantidos fixos nesta etapa. Os intervalos de valores testados nesta rodada estão apresentados a seguir, com os valores selecionados pela segunda rodada destacados em negrito.

RandomForest

- Total de $2 \times 3 \times 3 \times 1 \times 1 = 18$ combinações ($18 \times$ 5-fold $= 90$ modelos treinados)
  - *n_estimators*: [**250**, 300]
  - *max_depth*: [None, **20**, 30]
  - *min_samples_split*: [3, **4**, 5]
  - *min_samples_leaf*: [**2**]
  - *max_features*: [**'log2'**]

Os resultados obtidos na terceira rodada, bem como a comparação com as rodadas anteriores, a partir da avaliação pelas métricas de regressão, estão apresentados na Tabela 5. Os valores dos hiperparâmetros selecionados ao final dessa rodada também são descritos a seguir.

RandomForest

- *n_estimators*: 300
- *max_depth*: 20
- *min_samples_split*: 3
- *min_samples_leaf*: 2
- *max_features*: 'log2'

Tabela 5 – Desempenho comparativo dos modelos após a terceira rodada de otimização

Modelo	R^2	**MAE**	**MSE**	**RMSE**	**SMAPE**
1º	0,8152	**0,1107**	0,0352	0,1876	**84,39%**
2º	0,8165	0,1108	**0,0349**	**0,1869**	87,65%
3º	**0,8166**	**0,1107**	0,0350	0,1870	87,63%
Δ% (entre 2º e 3º)	0,01%	0,09%	0,29%	0,05%	0,02%

Algumas observações finais podem ser feitas a partir dessa última validação. As variações entre a segunda e terceira rodada foram mínimas, com diferenças percentuais inferiores a 0,50% para todas as métricas analisadas. Além disso, observou-se uma leve regressão em duas métricas, ainda que pouco significativa. Dado o elevado custo computacional envolvido e a tendência decrescente dos ganhos obtidos com cada rodada, decidiu-se

encerrar a busca exaustiva por combinações de hiperparâmetros e adotar o modelo resultante da terceira rodada como versão final.

É importante destacar que, embora este estudo tenha se baseado na replicação do índice de risco de fogo utilizando apenas cinco variáveis (temperatura do ar, umidade relativa, precipitação acumulada, número de dias consecutivos sem chuva e concentração atmosférica de fumaça), o cálculo oficial do índice considera um conjunto mais amplo de fatores. Entre eles, incluem-se a elevação topográfica e o tipo da vegetação, por exemplo (SETZER; SISMANOGLU; SANTOS, 2019).

4.3 Avaliação dos Resultados

A partir do modelo final definido, realizaram-se diferentes análises para investigar seu comportamento e avaliar sua robustez. Essas análises incluíram a mensuração da importância das variáveis, o estudo dos resíduos, a aplicação de testes de sensibilidade e a comparação espacial entre os valores previstos e observados. Cada uma dessas abordagens contribui para a interpretação crítica dos resultados, conforme descrito nas subseções seguintes.

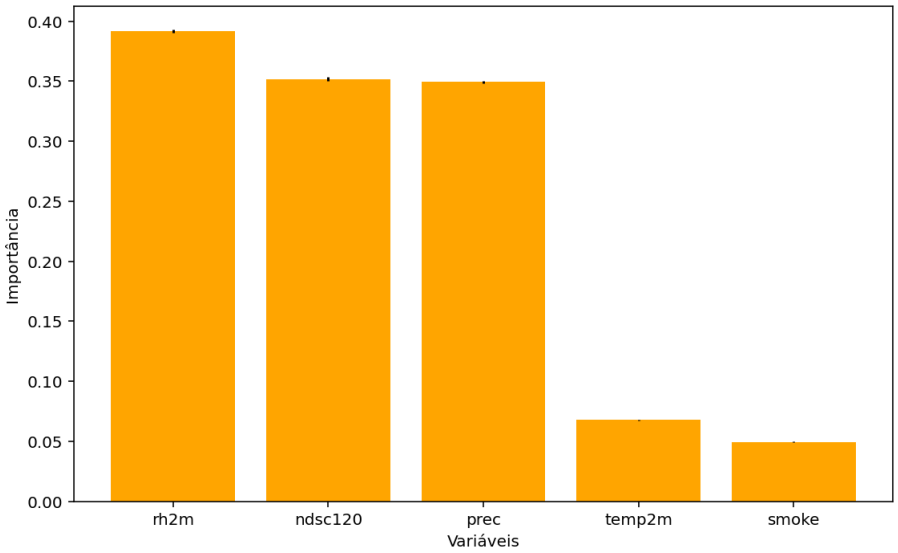
4.3.1 Análise da Importância das Variáveis

Para compreender o impacto de cada variável na construção do modelo final, realizou-se, inicialmente, uma análise de importância por permutação. Essa abordagem avalia como a capacidade preditiva do modelo é afetada ao se embaralhar os valores de uma variável de entrada, sem alterar as demais. Quanto maior for a queda no desempenho, mais relevante é considerada a variável permutada. O método foi aplicado sobre o conjunto de teste, garantindo que os resultados refletissem a capacidade real de generalização do modelo.

A técnica foi implementada por meio da função *permutation_importance*, disponibilizada pela biblioteca *scikit-learn*. O modelo é utilizado para realizar novas predições com base no conjunto modificado a partir da permutação, e a perda no desempenho (em relação ao conjunto original) é registrada. A métrica utilizada para quantificar a perda do desempenho foi o coeficiente de determinação (R^2), pelos mesmos motivos já discutidos anteriormente, quando adotado como critério principal em *GridSearchCV*. Uma importância de 0,35, por exemplo, indica que embaralhar os valores dessa variável reduz, em média, 0,35 unidade de R^2 do modelo. Para cada variável, foram realizadas 10 permutações aleatórias e calculou-se a média e o desvio padrão das perdas observadas.

Os valores obtidos foram organizados em ordem decrescente de importância média e apresentados na Figura 10. A Tabela 6 apresenta os resultados completos, incluindo as médias e desvios padrão das importâncias por permutação. Os nomes das variáveis seguem a nomenclatura utilizada nos conjuntos de dados disponibilizados pelo INPE: *temp2m* (temperatura do ar), *ndsc120* (número de dias consecutivos sem chuva nos últimos 120 dias), *rh2m* (umidade relativa do ar), *prec* (precipitação acumulada) e *smoke* (concentração atmosférica de fumaça).

Figura 10 – Importância das variáveis por permutação baseada na perda média de R^2



Fonte: Autoria própria

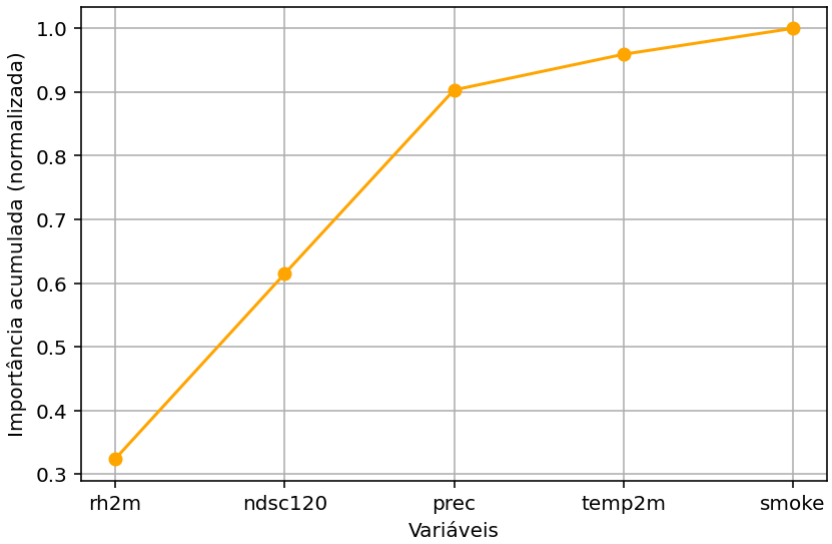
Tabela 6 – Médias e desvios padrão da importância por permutação das variáveis

Variável	**Média da perda (importância)**	**Desvio padrão**
Umidade relativa (rh2m)	0,391656	0,001225
N.º dias s/ chuva (ndsc120)	0,351764	0,001653
Precipitação acumulada (prec)	0,349289	0,000938
Temperatura do ar (temp2m)	0.067776	0.000372
Concentração atm. de fumaça (smoke)	0.049262	0.000368

Além da análise individual, foi realizada a normalização da importância relativa de cada variável em relação à soma das importâncias. A Figura 11 apresenta a importância acumulada das variáveis. Essa visualização permite identificar quais variáveis, ao serem incluídas, impactam significativamente o desempenho do modelo. A partir dela,

observa-se que as três variáveis mais importantes — umidade relativa, número de dias sem chuva e precipitação acumulada — representam um impacto de aproximadamente 90% na capacidade preditiva do modelo.

Figura 11 – Importância acumulada das variáveis após normalização

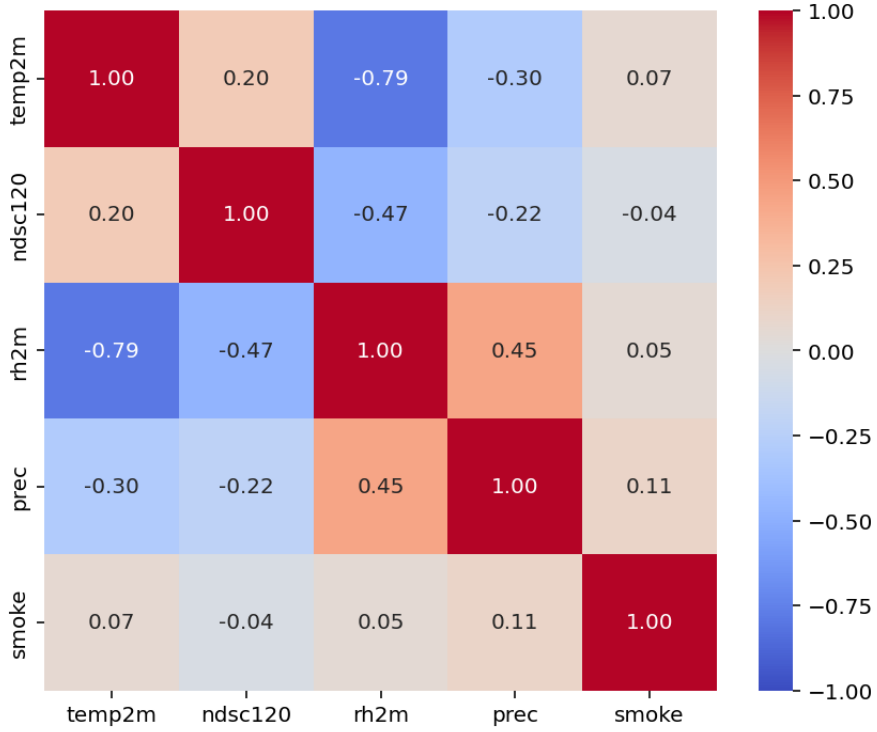


Fonte: Autoria própria

Também foi elaborada uma matriz de correlação entre as variáveis de entrada utilizadas na modelagem. Essa matriz permite verificar a existência de relações lineares entre duas variáveis, evidenciando colinearidades relevantes na interpretação e desempenho do modelo. Sua construção baseou-se em funções disponibilizadas pela biblioteca *pandas*, que calcula o coeficiente de correlação entre todas as combinações possíveis de variáveis numéricas contínuas.

A Figura 12 apresenta a matriz de correlação obtida com base no coeficiente de Pearson. Essa medida é amplamente adotada em aplicações de aprendizado de máquina para avaliar a dependência linear entre variáveis numéricas, sendo adequada às variáveis meteorológicas e ambientais analisadas neste trabalho (HASTIE; TIBSHIRANI; FRIEDMAN, 2009). Valores próximos de $+1$ indicam uma forte correlação linear positiva, ou seja, variáveis que tendem a crescer juntas. Valores próximos de -1 indicam correlação negativa forte, onde o aumento de uma variável está associado à diminuição de outra. Por fim, valores próximos de 0 indicam ausência de relação linear entre as variáveis analisadas.

Figura 12 – Matriz de correlação linear baseadas no coeficiente de Pearson

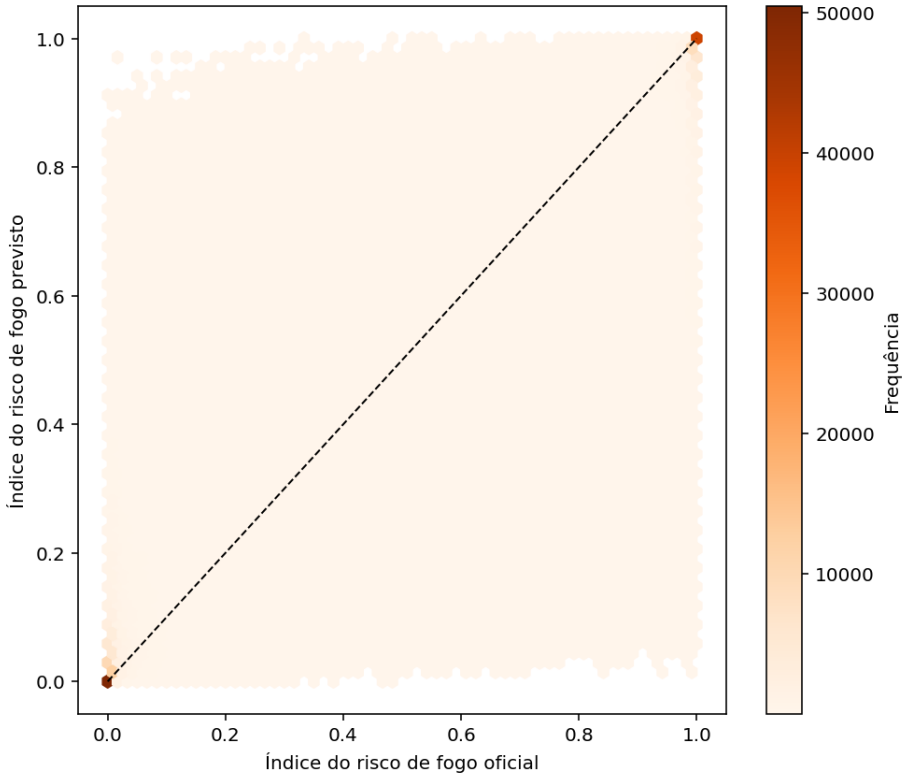


Fonte: Autoria própria

4.3.2 Análise Residual

Para avaliar a qualidade do modelo ajustado e identificar possíveis padrões não capturados pelas predições, foi realizada uma análise residual com base no conjunto de teste. Resíduos são definidos pela diferença entre os valores reais (y) e os valores previstos pelo modelo ($\hat{y}$), representando o erro de predição em cada observação. Hastie, Tibshirani e Friedman (2009) destacam que a análise dos resíduos permite a identificação de padrões nos erros, pelo viés sistemático (quando o modelo tende a superestimar ou subestimar em certas faixas de valores) ou por variações na dispersão dos erros. Esses comportamentos não são facilmente perceptíveis utilizando apenas as métricas agregadas avaliadas anteriormente. A Figura 13 apresenta um gráfico de dispersão por densidade (*hexbin*) entre esses dois conjuntos. A linha diagonal tracejada representa a predição ideal ($y = \hat{y}$), enquanto as regiões mais densas indicam maior concentração de observações.

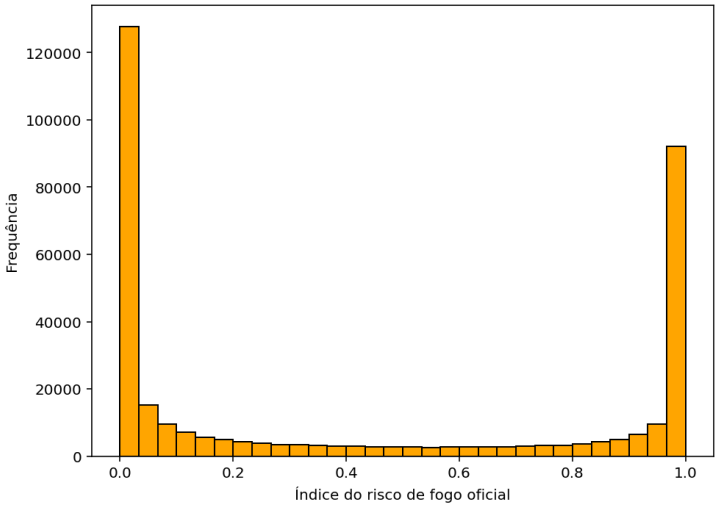
Figura 13 – Dispersão por densidade entre valores reais e previstos do índice do risco de fogo



Fonte: Autoria própria

Dois aspectos se destacam no gráfico acima: (1) a dispersão dos dados em todo o gráfico, resultado do elevado número de observações na amostra de teste (346.352 pontos); e (2) a forte concentração de pontos próximos às extremidades, onde ambos os índices do risco de fogo oficial e previsto assumem valores 0 ou 1. Esse comportamento reflete a própria distribuição empírica do índice oficial, evidenciado pelo histograma da Figura 14, que ilustra a frequência dos valores reais do índice.

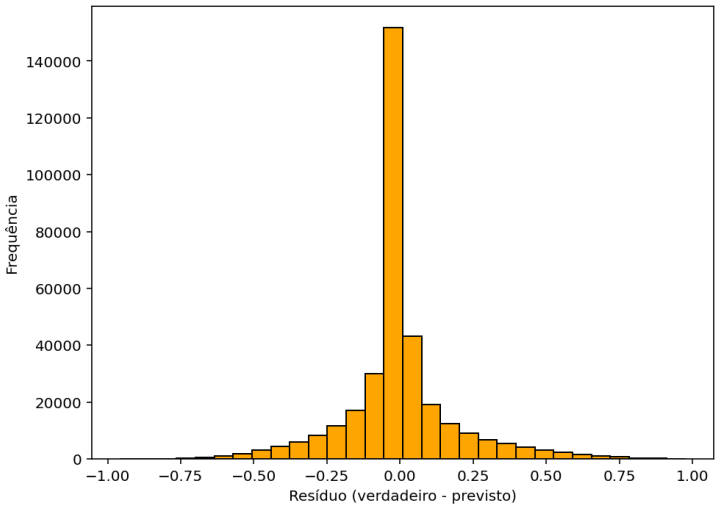
Figura 14 – Distribuição dos valores oficiais do índice do risco de fogo no conjunto de teste



Fonte: Autoria própria

Retornando à análise residual, a Figura 15 apresenta um histograma dos resíduos, permitindo avaliar a distribuição dos erros cometidos pelo modelo. Observa-se que os resíduos estão centrados em torno de zero, apresentados com forma similar a uma distribuição normal. Há uma leve assimetria à esquerda, indicando uma tendência a erros negativos — i.e., casos em que o modelo superestima o índice do risco de fogo em relação ao valor real. Destaca-se ainda a ausência de *outliers* e de regiões com padrões sistemáticos de erro, sugerindo um bom comportamento do modelo em relação à acurácia.

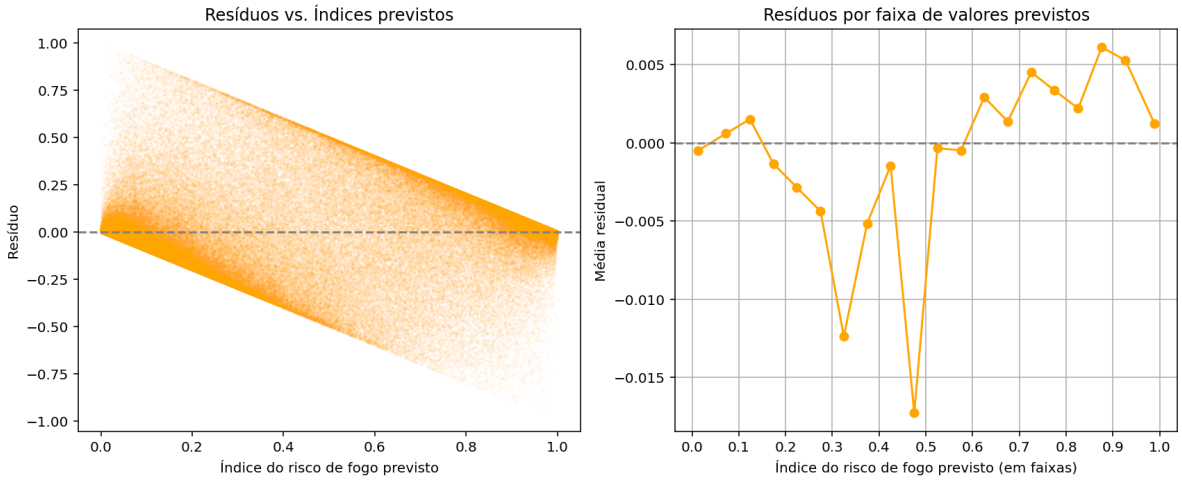
Figura 15 – Distribuição dos valores residuais obtidos através do modelo



Fonte: Autoria própria

A Figura 16 apresenta dois gráficos complementares. No painel esquerdo, exibe-se um gráfico de dispersão dos resíduos em função dos valores previstos do índice de risco de fogo. Observa-se uma concentração maior de pontos em torno da linha zero, com distribuição simétrica. O aprofundamento da análise requer as informações exibidas pelo painel esquerdo, que apresenta a média dos resíduos agrupados em 20 faixas (*bins*) de valores previstos. Essa visualização permite identificar variações sistemáticas localizadas, como viés em predições muito altas ou muito baixas. O modelo apresenta médias residuais próximas de zero em praticamente todo o domínio, com maior variação negativa entre os intervalos de 0,3 e 0,5, ainda assim de baixa magnitude.

Figura 16 – Resíduos e médias por faixa em função do índice previsto do risco de fogo



Fonte: Autoria própria

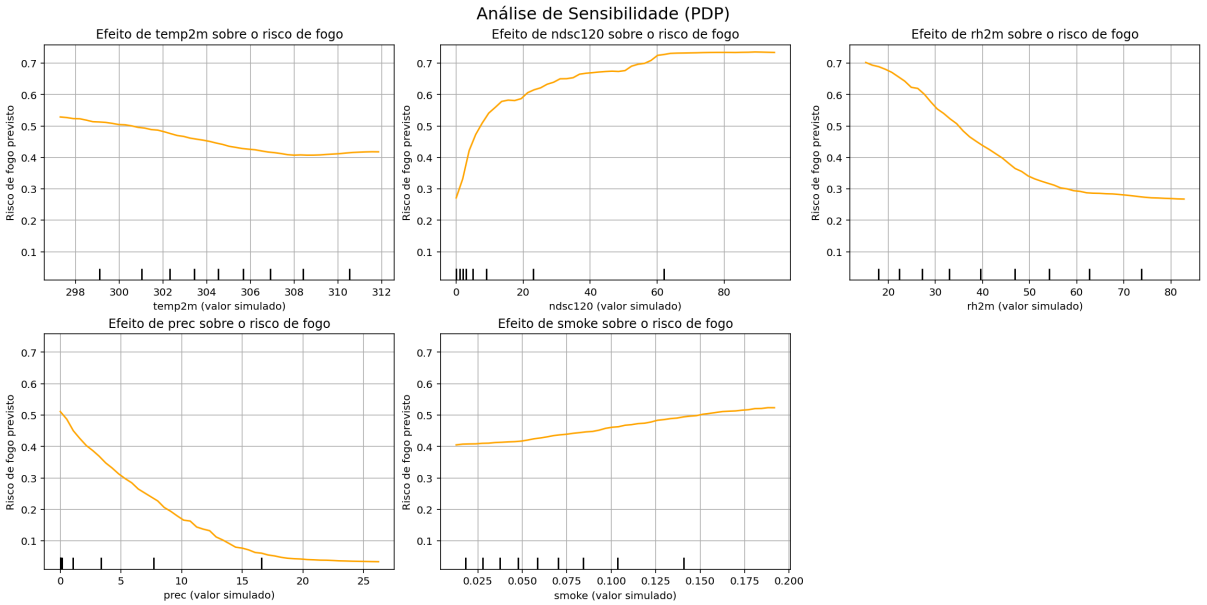
4.3.3 Testes de Sensibilidade

Adicionalmente, realizaram-se testes de sensibilidade visando compreender como as variáveis de entrada afetam individualmente as predições do modelo. A análise desses testes permite observar o comportamento da previsão de saída do modelo à medida que os valores de uma variável específica são alterados, mantendo-se as demais constantes.

Para aplicação desse método, adotou-se a técnica de *Partial Dependence Plots* (PDP), implementada também pela biblioteca *scikit-learn* por meio da classe *PartialDependenceDisplay*. O PDP calcula a média das predições do modelo a partir da variação dos valores de uma variável, preservando a distribuição original das demais variáveis conforme o conjunto de dados de treinamento. A resposta média do modelo é definida em função dos valores simulados a partir da variável analisada. A Figura 17 apresenta os gráficos de

dependência parcial das cinco variáveis de entrada, evidenciando o efeito isolado de cada uma na predição do risco de fogo conforme variação numérica.

Figura 17 – Testes de sensibilidade aplicados sobre as cinco variáveis de entrada



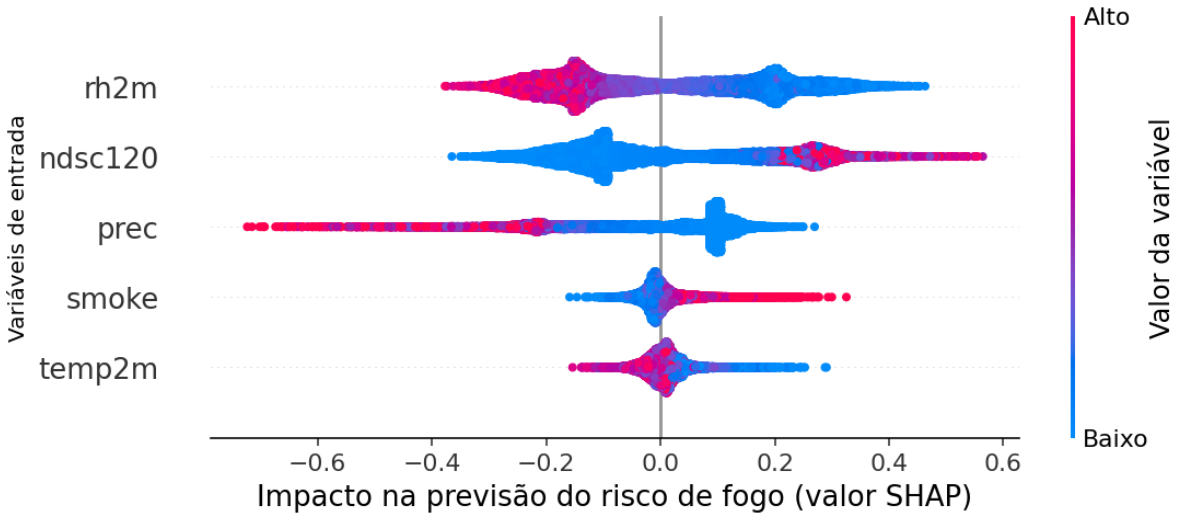
Fonte: Autoria própria

Além dos testes de sensibilidade por dependência parcial, utilizou-se também a abordagem SHAP (***SH****apley* ***A****dditive ex****P****lanations*). O objetivo do método é atribuir, para cada variável de entrada, um valor que represente sua contribuição individual para a predição realizada pelo modelo em cada instância (i.e., cada amostra individual do conjunto de dados). Por exemplo, caso o modelo estime um risco de fogo igual a 0,78, o SHAP calcula o quanto dessa predição é atribuível à temperatura, o quanto à umidade, e assim por diante. Neste trabalho, adotou-se o algoritmo *TreeExplainer*, disponibilizado pela biblioteca *shap*, otimizado para modelos baseados em árvores, como o RF.

A Figura 18 apresenta um gráfico de resumo dos valores SHAP, que permite a visualização simultânea de diversas informações. No eixo vertical, encontram-se as variáveis de entrada, ordenadas pela média dos valores absolutos dos SHAPs — medida relacionada à importância média de cada variável nas predições. No eixo horizontal, cada ponto corresponde ao valor SHAP atribuído a uma instância específica da base de dados, indicando o quanto aquela variável contribuiu para aumentar ou reduzir a predição em relação ao valor médio predito pelo modelo (i.e., a predição que seria feita na ausência de qualquer informação de entrada). A coloração dos pontos representa os valores originais da variável para cada instância: pontos em vermelho indicam valores altos da variável; pontos em azul

representam valores baixos. A distribuição e concentração dos pontos também permitem a observação da densidade e da consistência das predições ao longo do domínio da variável.

Figura 18 – Importância e influência das variáveis segundo os valores SHAP



Fonte: Autoria própria

4.3.4 Análise Espacial Comparativa

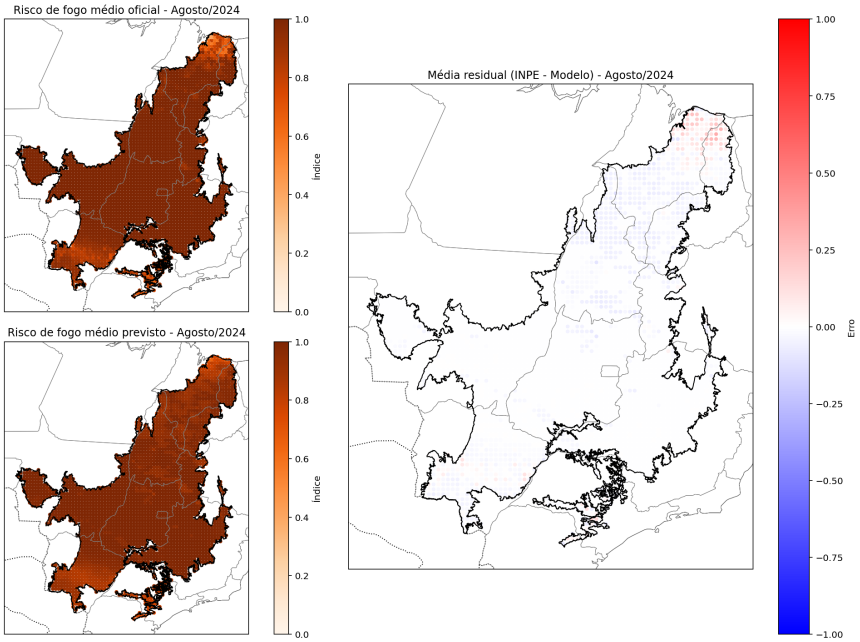
Complementando a avaliação do modelo e verificando sua capacidade de capturar padrões espaciais relevantes, realizou-se uma comparação geográfica entre os valores reais e previstos do índice de risco de fogo. Essa abordagem permite não apenas avaliar o desempenho numérico do modelo, mas também inspecionar a consistência das previsões em relação ao comportamento observado do fenômeno.

Para analisar a robustez do modelo sob diferentes condições sazonais, foram selecionados três recortes temporais distintos: agosto de 2024, caracterizado pelo pico da estação seca e associado a elevados índices de área queimada (PEIXOTO, 2024); janeiro de 2025, inserido na estação chuvosa, período no qual o índice de risco de fogo tende a ser reduzido (PRIZIBISCZKI, 2022); e maio de 2025, adotado como um terceiro cenário neutro, sem a predominância de fenômenos meteorológicos específicos e representando o último mês disponível na base de dados. Essa seleção permite verificar a capacidade do modelo em reproduzir padrões espaciais sob diferentes regimes climáticos.

Para cada um desses meses, calculou-se a média mensal do índice de risco de fogo fornecido pelo INPE e das respectivas previsões do modelo, considerando todos os pontos geográficos do bioma Cerrado. A diferença entre essas médias resulta na média residual de cada ponto. Os resultados são apresentados por meio de três mapas por mês: o primeiro

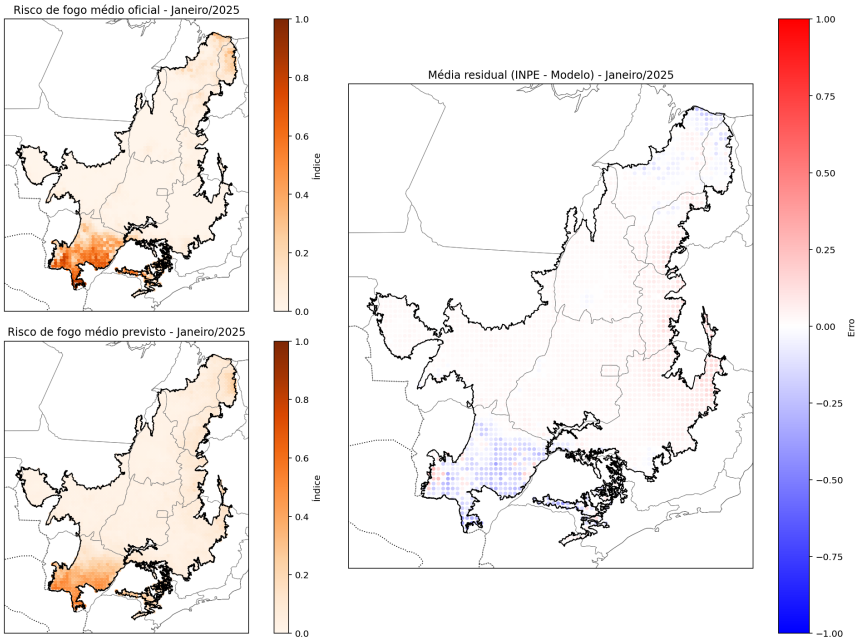
mostra o índice médio do risco de fogo segundo os dados reais; o segundo, os valores previstos pelo modelo; e o terceiro (à direita) exibe a média dos resíduos espaciais, obtidos pela subtração entre os dois anteriores. As Figuras 19, 20 e 21 apresentam essa comparação para os meses de agosto (2024), janeiro (2025) e maio (2025), permitindo a identificação de regiões onde o modelo superestima ou subestima os valores observados.

Figura 19 – Comparação espacial entre o índice de risco de fogo médio real, previsto e média residual (Agosto/2024)



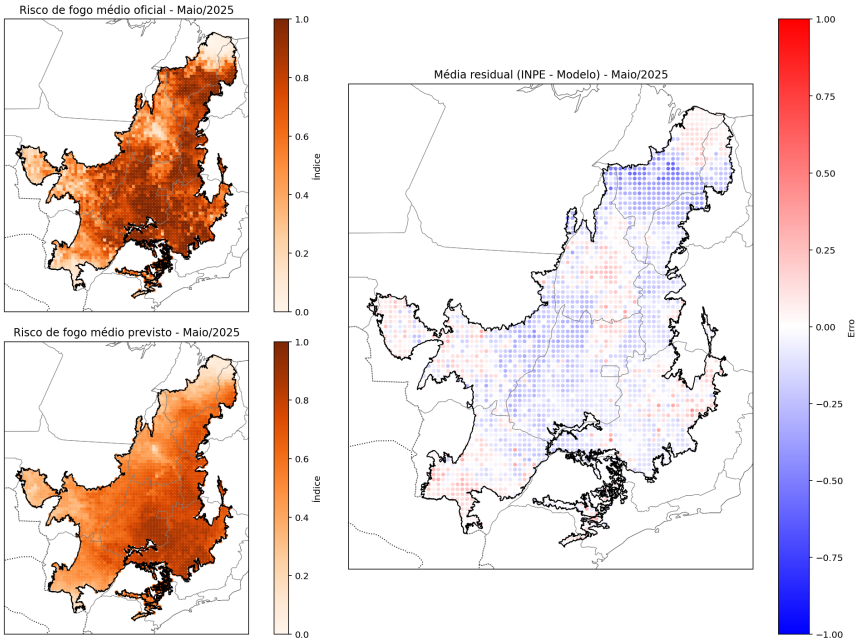
Fonte: Autoria própria

Figura 20 – Comparação espacial entre o índice de risco de fogo médio real, previsto e média residual (Janeiro/2025)



Fonte: Autoria própria

Figura 21 – Comparação espacial entre o índice de risco de fogo médio real, previsto e média residual (Maio/2025)



Fonte: Autoria própria

5 Resultados e Discussão

O desenvolvimento do trabalho teve como objetivo inicial a replicação do índice do risco de fogo do INPE a partir de cinco variáveis ambientais e meteorológicas – temperatura do ar, precipitação acumulada, número de dias sem chuva, umidade relativa e concentração atmosférica de fumaça – utilizando dados abertos disponibilizados pelo Programa Queimadas. O foco foi direcionado à análise do impacto dessas variáveis no bioma Cerrado, utilizando dados disponibilizados entre agosto de 2023 e maio de 2025.

Os dados foram previamente tratados para garantir que todas as variáveis apresentassem valores válidos, por meio de processos de limpeza, interpolação e normalização, quando necessário. Foram avaliados três modelos de aprendizado de máquina com abordagens distintas: o *RandomForest*, baseado na técnica de *bagging*; o *XGBoost*, fundamentado na técnica de *boosting*; e o *MLPRegressor*, modelo baseado em redes neurais artificiais do tipo MLP. Para cada técnica, desenvolveram-se modelos com variações de hiperparâmetros, utilizando cinco divisões de validação cruzada combinada com busca exaustiva. Para fins de modelagem, 80% do conjunto de dados foi destinado ao treinamento e os 20% restantes à etapa de teste. Foi selecionada, ao final, a configuração que obteve o melhor desempenho segundo o coeficiente de determinação (R^2), adotado como métrica principal por quantificar a proporção da variância explicada pelo modelo e por seu uso recorrente em problemas de regressão.

O modelo baseado em RF apresentou o melhor desempenho em todas as métricas consideradas neste trabalho (R^2, MAE, MSE, RMSE, SMAPE). Especificamente, no que se refere ao R^2, o modelo baseado em árvores aleatórias apresentou um valor de 0,8152, em comparação com 0,8121 e 0,8082, obtidos por *XGBoost* e *MLPRegressor*, respectivamente. Esse resultado era esperado, conforme discutido previamente no Capítulo 3, e está alinhado com as conclusões de trabalhos similares, como Rubí e Gondim (2024), Gholamnia et al. (2020), Galizia e Rodrigues (2019) e Oliveira et al. (2022), que também apontaram o RF como o modelo preditivo de melhor desempenho em aplicações similares relacionadas a incêndios florestais. O algoritmo apresentou robustez e boa capacidade de generalização do modelo em cenários com variáveis ambientais ruidosas e heterogêneas.

Na sequência, realizou-se o processo de otimização do modelo, observando-se que o ajuste de hiperparâmetros teve impacto limitado no desempenho final do modelo. As variações nas métricas de avaliação entre os diferentes modelos de RF foram inferiores a 1% na segunda rodada e menores que 0,5% na terceira, indicando um ponto de estabilidade no processo de treinamento. Considerando que o foco deste trabalho está na análise

da influência das variáveis e não na maximização absoluta do desempenho preditivo do índice de risco de fogo, além do alto custo computacional envolvido em novas rodadas de validação, optou-se por encerrar a busca exaustiva após a terceira rodada, com um modelo final cujo valor de R^2 foi de 0,8166.

A partir da definição do modelo final, realizaram-se análises quantitativas visando melhor entender suas correlações e os fatores que mais influenciam suas predições. O primeiro estudo consistiu na análise de importância das variáveis por permutação, técnica que permite mensurar a influência preditiva isolada de cada variável. Conforme ilustrado na Figura 10, a variável de maior impacto foi a umidade relativa, cuja permutação causou uma queda de 0,392, com desvio-padrão de 0,001, a maior redução observada. Em seguida, destaca-se o número de dias sem chuva, a precipitação acumulada, a temperatura do ar e, por último, a concentração atmosférica de fumaça. As três primeiras variáveis, em conjunto, foram responsáveis por cerca de 90% da capacidade explicativa do modelo, evidenciando sua predominância na determinação do índice de risco de fogo.

A matriz de correlação linear apresentada na Figura 12 reforça algumas relações esperadas entre as variáveis de entrada. Observa-se uma correlação negativa significativa entre precipitação acumulada e temperatura, sugerindo que dias mais chuvosos tendem a apresentar temperaturas mais baixas. Também foi identificada uma correlação positiva entre umidade relativa e precipitação, além de uma correlação negativa acentuada entre umidade relativa e número de dias sem chuva — comportamento coerente com o efeito de secas prolongadas na umidade do ar. A concentração de fumaça, por outro lado, apresentou fracas correlações com as demais variáveis, indicando que sua variabilidade pode estar associada a fatores externos não identificados.

Em seguida, realizou-se a análise residual do modelo, visando avaliar possíveis padrões nos erros de predição. Os resíduos foram calculados pela diferença entre os valores reais e os valores previstos pelo modelo, representando, assim, o erro individual de cada observação. A Figura 13 apresenta a relação entre os valores reais e previstos, onde se observa uma elevada concentração de observações nas extremidades, em regiões próximas a 0 e 1. Esse comportamento reflete a própria natureza da distribuição do índice de risco de fogo na base de dados original, e não uma falha do modelo em si, observável na Figura 14.

A Figura 15 apresenta a distribuição dos resíduos em forma de histograma. Observa-se uma curva aproximadamente simétrica, próxima à distribuição normal, com leve assimetria à esquerda. Esse comportamento sugere uma tendência sutil do modelo a superestimar o índice real de risco de fogo. Ainda assim, a centralização em torno de zero e a ausência de caudas longas indicam não haver presença significativa de *outliers* ou erros sistemáticos.

Por fim, a Figura 16 complementa a análise ao detalhar o comportamento dos resíduos em função dos valores previstos. No gráfico da direita, que apresenta as médias dos resíduos em 20 faixas (*bins*) de valores previstos, observa-se que, mesmo nos piores casos, a média dos resíduos ultrapassa ligeiramente -0,015. A maioria das faixas mantém médias residuais próximas de zero, evidenciando que o modelo mantém erros baixos e consistentes em diferentes níveis do índice previsto. Esse comportamento reforça a robustez do modelo ao longo de todo o domínio de predição, sem viés observado em faixas específicas, reforçando sua capacidade de generalização.

Após a análise residual, foram conduzidos testes de sensibilidade visando avaliar como as alterações nas variáveis de entrada impactam as predições do modelo. A Figura 17 apresenta os resultados obtidos por meio dos gráficos de dependência parcial (PDP), que simulam variações individuais nos valores de cada variável, mantendo as demais constantes e observando a resposta média do modelo.

Alguns comportamentos esperados foram confirmados. O aumento no número de dias sem chuva provocou um crescimento expressivo na predição do risco de fogo, variando seu índice de aproximadamente 0,3 até 0,7. Da mesma forma, a elevação da precipitação acumulada resultou na redução da previsão do risco, que caiu de 0,5 para valores próximos de zero. A umidade relativa também apresentou comportamento esperado: valores mais altos dessa variável levaram à redução do risco estimado, de cerca de 0,7 para 0,3. Já o aumento da concentração atmosférica de fumaça causou apenas uma leve elevação no risco previsto, com uma diferença em torno de 0,1.

Por outro lado, observou-se um comportamento não intuitivo em relação à temperatura do ar: o aumento da variável não resultou em um crescimento proporcional no risco de fogo — o efeito foi inverso. Esse comportamento contradiz parte das evidências discutidas por Setzer, Sismanoglu e Santos (2019), que apontam o aumento da temperatura (especialmente acima de 30 °C) como um dos fatores de intensificação do risco. Apesar disso, o mesmo estudo menciona a forte influência do número de "Dias de Secura" (número de dias seguidos sem nenhuma precipitação durante os últimos 120 dias) e da umidade relativa no agravamento do risco de incêndios, o que foi corroborado pelos resultados obtidos neste trabalho.

Além dos testes de sensibilidade por dependência parcial, aplicou-se também a abordagem SHAP, que quantifica a contribuição individual de cada variável de entrada para a predição do modelo em cada instância. Os resultados dos valores SHAP observados pela Figura 18 corroboram com as observações anteriores: valores baixos de umidade relativa, precipitação acumulada e número de dias sem chuva apresentaram os maiores impactos no risco de fogo previsto. Já a concentração de fumaça e a temperatura do ar mostraram influência reduzida e menor variação nos valores SHAP — reforçando seu papel

secundário na predição do modelo, conforme já observado nos gráficos de dependência parcial.

Por fim, realizou-se uma comparação espacial entre os valores do índice de risco de fogo oficiais, as predições do modelo e seus respectivos resíduos médios. O objetivo foi identificar padrões regionais de acerto e erro do modelo, além de possíveis inconsistências sistemáticas em áreas específicas do bioma Cerrado. Para o mês de agosto de 2024 — período de maior registro de queimadas no Cerrado — observou-se uma taxa de erro mínima entre os valores reais e previstos. A maioria do bioma apresentou resíduos próximos de zero, com uma leve tendência à subestimação do risco no extremo norte da região.

Em janeiro de 2025, caracterizado por ser parte de uma estação chuvosa e, portanto, menos suscetível a queimadas, o modelo novamente demonstrou boa acurácia. Os resíduos médios permaneceram baixos, com uma leve superestimação no sudoeste do bioma. O mês de maio de 2025, que apresentou níveis intermediários de risco, revelou um padrão distinto: observou-se maior dispersão dos resíduos em todo o Cerrado, com erros de média-baixa magnitude em diferentes regiões. As discrepâncias mais notáveis ocorreram no norte do bioma, onde o modelo apresentou maiores superestimações. Esse comportamento pode estar relacionado a mudanças transitórias nas condições climáticas típicas do período ou à menor previsibilidade intrínseca desse cenário intermediário. Isso evidencia a robustez espacial do modelo nos extremos da sazonalidade (seco e chuvoso), mas aponta para uma maior instabilidade em meses de transição climática. Essa limitação deve ser considerada em aplicações operacionais que dependam de previsões mais precisas em períodos de instabilidade climática.

6 Considerações finais

Este trabalho teve como objetivo replicar o índice de risco de fogo do INPE utilizando variáveis ambientais e meteorológicas acessíveis, com foco no bioma Cerrado. Foram aplicadas técnicas de aprendizado de máquina para estimar esse índice e investigar a influência relativa de cada variável no comportamento preditivo dos modelos.

As análises indicaram que três variáveis foram determinantes para a capacidade preditiva do modelo: umidade relativa, número de dias sem chuva e precipitação acumulada. A forte influência dessas variáveis reforça o papel determinante das condições de estiagem prolongada como fator de risco. Similarmente, Oliveira et al. (2022) evidenciam que as variáveis climáticas explicam 56% da variância em seu modelo, similarmente as observações feitas neste estudo. Os testes de sensibilidade confirmaram esse padrão, demonstrando a robustez das predições em diferentes cenários. Conclui-se, portanto, que períodos com baixa umidade, ausência prolongada de chuvas e precipitação reduzida devem acionar mecanismos de alerta por parte das autoridades ambientais, dada a maior probabilidade de ocorrência de incêndios. Estes, que embora façam parte da dinâmica natural do Cerrado, devem permanecer sob controle das autoridades para que não se tornem um mega incêndio (DURIGAN, 2020; FIDELIS et al., 2018).

Apesar dos resultados expressivos, o trabalho apresenta limitações significativas. O conjunto de variáveis utilizadas restringiu-se a cinco métricas principais, não incluindo fatores relevantes como velocidade e direção do vento, radiação solar, pressão atmosférica, elevação topográfica ou concentração de diversos gases observados na atmosfera, cuja ausência pode limitar a capacidade de generalização do modelo. Além disso, os dados diretos sobre os impactos das queimadas — como área queimada — estão disponíveis apenas em resoluções mensais, dificultando a avaliação direta da consequência dos riscos previstos. Também não foram consideradas variáveis sobre à cobertura vegetal ou ao histórico anual de queimadas, informações que podem indicar padrões sazonais e contribuir para uma previsão mais precisa e contextualizada.

Para estudos futuros, recomenda-se a ampliação do conjunto de variáveis consideradas, juntamente da integração dos dados de cobertura do solo, uso de índices relacionados à topografia e biomassa, além da correlação direta entre o risco de fogo previsto com o impacto observado pelas áreas queimadas. Por exemplo, Wu et al. (2022) demonstraram que a altitude foi o fator mais influente na propagação de fogo em Heilongjiang, China – variável ausente neste estudo. Tais aprimoramentos podem contribuir na construção de

modelos mais robustos e potencialmente superiores ao próprio índice de fogo atualmente disponibilizado pelo INPE.

Referências

AGGARWAL, C. C. *Neural Networks and Deep Learning: A Textbook*. Springer, 2018. ISBN 978-3-319-94462-3. Disponível em: <https://doi.org/10.1007/978-3-319-94463-0>.

Anaconda Inc. *Anaconda Distribution*. 2025. <https://www.anaconda.com>. Acesso em: 23 mar. 2025.

ARAGãO, L. E. O. C.; ANDERSON, L. O.; FONSECA, M. G.; ROSAN, T. M.; VEDOVATO, L. B.; WAGNER, F. H.; SILVA, C. V. J.; JUNIOR, C. H. L. S.; ARAI, E.; AGUIAR, A. P.; BARLOW, J.; BERENGUER, E.; DEETER, M. N.; DOMINGUES, L. G.; GATTI, L.; GLOOR, M.; MALHI, Y.; MARENGO, J. A.; MILLER, J. B.; PHILLIPS, O. L.; SAATCHI, S. 21st century drought-related fires counteract the decline of amazon deforestation carbon emissions. *Nature Communications*, England, v. 9, n. 1, p. 536, Feb 2018.

AYYADEVARA, V. K. Random forest. In: *Pro Machine Learning Algorithms*. Apress, 2018. p. 105–116. Disponível em: <http://dx.doi.org/10.1007/978-1-4842-3564-5_5>.

BREIMAN, L. Random forests. *Machine Learning*, v. 45, n. 1, p. 5–32, 2001.

BURDEN, R.; FAIRES, J. *Numerical Analysis*. Brooks/Cole, Cengage Learning, 2011. ISBN 9780538735643. Disponível em: <https://books.google.com.br/books?id=KlfrjCDayHwC>.

CANADIAN FOREST FIRE WEATHER INDEX. 2025. Disponível em: <https://cwfis.cfs.nrcan.gc.ca/maps/fw?type=fwi>. Acesso em: 25 mai. 2025.

CHEN, T.; GUESTRIN, C. Xgboost: A scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY, USA: ACM, 2016. p. 785–794.

CNN. Queimadas provocaram prejuízos econômicos de r$ 1,3 bilhão, diz cnm. *CNN Brasil*, set. 2024. Disponível em: <https://www.cnnbrasil.com.br/economia/macroeconomia/queimadas-provocaram-prejuizos-economicos-de-r-13-bilhao-diz-cnm/>.

COLLI, G. R.; VIEIRA, C. R.; DIANESE, J. C. Biodiversity and conservation of the cerrado: recent advances and old challenges. *Biodiversity and Conservation*, v. 29, n. 5, p. 1465–1475, Apr 2020. ISSN 1572-9710. Disponível em: <https://doi.org/10.1007/s10531-020-01967-x>.

CORTEZ, P.; MORAIS, A. d. J. R. *A data mining approach to predict forest fires using meteorological data*. 2007.

COUTINHO, E. R.; SILVA, R. M.; DELGADO, A. R. S. Utilização de técnicas de inteligência computacional na predição de dados meteorológicos. *Revista Brasileira de Meteorologia*, Sociedade Brasileira de Meteorologia, v. 31, n. 1, p. 24–36, Jan 2016. ISSN 0102-7786. Disponível em: <https://doi.org/10.1590/0102-778620140115>.

DROUGHT AND FIRE OBSERVATORY AND EARLY WARNING SYSTEM. 2025. Disponível em: <http://disarmfire.eu/>. Acesso em: 25 mai. 2025.

DURIGAN, G. Zero-fire: Not possible nor desirable in the cerrado of brazil. *Flora*, v. 268, p. 151612, 2020. ISSN 0367-2530. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0367253020300761>.

EUROPEAN FOREST FIRE INFORMATION SYSTEM. 2025. Disponível em: <https://forest-fire.emergency.copernicus.eu/>. Acesso em: 25 mai. 2025.

FIDELIS, A.; ALVARADO, S. T.; BARRADAS, A. C. S.; PIVELLO, V. R. The year 2017: Megafires and management in the cerrado. *Fire*, v. 1, n. 3, p. 49, 2018. ISSN 2571-6255. Disponível em: <https://www.mdpi.com/2571-6255/1/3/49>.

FIRE INFORMATION FOR RESOURCE MANAGEMENT SYSTEM US / CANADA. 2025. Disponível em: <https://firms2.modaps.eosdis.nasa.gov/usfs/>. Acesso em: 25 mai. 2025.

GALIZIA, L. F. d. C.; RODRIGUES, M. Modeling the influence of eucalypt plantation on wildfire occurrence in the brazilian savanna biome. *Forests*, v. 10, n. 10, p. 844, 2019. ISSN 1999-4907. Disponível em: <https://www.mdpi.com/1999-4907/10/10/844>.

GHOLAMNIA, K.; NACHAPPA, T. G.; GHORBANZADEH, O.; BLASCHKE, T. Comparisons of diverse machine learning approaches for wildfire susceptibility mapping. *Symmetry*, v. 12, n. 4, p. 604, 2020. ISSN 2073-8994. Disponível em: <https://www.mdpi.com/2073-8994/12/4/604>.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. *Deep Learning*. MIT Press, 2016. ISBN 9780262035613. Disponível em: <https://www.deeplearningbook.org>.

GUARALDO, L. Fogo queimou 88 milhões de hectares do cerrado nos últimos 39 anos. *IPAM Amazônia*, set. 2024. Disponível em: <https://ipam.org.br/fogo-queimou-88-milhoes-de-hectares-do-cerrado-nos-ultimos-39-anos/>.

HARPER, A. R.; DOERR, S. H.; SANTIN, C.; FROYD, C. A.; SINNADURAI, P. Prescribed fire and its impacts on ecosystem services in the uk. *Science of the Total Environment*, Elsevier, v. 624, p. 691–703, 2018.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd. ed. New York, NY: Springer, 2009. ISBN 978-0-387-84858-7. Disponível em: <https://link.springer.com/book/10.1007/978-0-387-84858-7>.

INSTITUTO NACIONAL DE PESQUISAS ESPACIAIS. *TerraBrasilis*. 2025. Disponível em: <https://terrabrasilis.dpi.inpe.br/>. Acesso em: 6 mai. 2025.

KOTSIANTIS, S. B.; ZAHARAKIS, I.; PINTELAS, P. Supervised machine learning: A review of classification techniques. In: *Emerging Artificial Intelligence Applications in Computer Engineering*. [S.l.: s.n.], 2007. v. 160, p. 3–24.

LEMOS, S. Queimadas no cerrado representam uma elevação na emissão de gases do efeito estufa. *Jornal da USP*, set. 2024. Disponível em: <https://jornal.usp.br/?p=806340>.

LU, G. Y.; WONG, D. W. An adaptive inverse-distance weighting spatial interpolation technique. *Computers & Geosciences*, v. 34, n. 9, p. 1044–1055, 2008. ISSN 0098-3004. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0098300408000721>.

MAIMON, O.; ROKACH, L. Data mining and knowledge discovery handbook. In: . 2nd. ed. [S.l.]: Springer Science+Business Media, LLC, 2010.

MARSLAND, S. *Machine Learning: An Algorithmic Perspective, Second Edition*. CRC Press, 2015. ISBN 9781498759786. Disponível em: <https://books.google.com.br/books?id=y_oYCwAAQBAJ>.

MCCULLOCH, W. S.; PITTS, W. A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics*, Springer, v. 5, n. 4, p. 115–133, 1943.

MITCHELL, T. M. *Machine Learning*. 1st. ed. New York: McGraw-Hill, 1997. ISBN 9780070428072.

MITTAL, A.; PATIDAR, S. Sentiment analysis on twitter data: A survey. In: *Proceedings of the 2019 7th International Conference on Computer and Communications Management (ICCCM)*. New York, NY, USA: ACM, 2019. p. 91–95. ISBN 978-1-4503-7195-7. Disponível em: <http://doi.acm.org/10.1145/3348445.3348466>.

NATURAL RESOURCES CANADA. *Canadian Forest Fire Weather Index (FWI) System*. 2025. Disponível em: <https://cwfis.cfs.nrcan.gc.ca/background/summary/fwi>. Acesso em: 01 jun. 2025.

OLIVEIRA, A.; SOARES-FILHO, B.; OLIVEIRA, U.; Van der Hoff, R.; CARVALHO-RIBEIRO, S.; OLIVEIRA, A.; SCHEEPERS, L.; VARGAS, B.; RAJãO, R. Costs and effectiveness of public and private fire management programs in the brazilian amazon and cerrado. *Forest Policy and Economics*, v. 127, p. 102447, 2021. ISSN 1389-9341. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1389934121000538>.

OLIVEIRA, U.; SOARES-FILHO, B.; BUSTAMANTE, M.; GOMES, L.; OMETTO, J. P.; RAJãO, R. Determinants of fire impact in the brazilian biomes. *Frontiers in Forests and Global Change*, v. 5, 2022. ISSN 2624-893X. Disponível em: <https://www.frontiersin.org/articles/10.3389/ffgc.2022.735017>.

ONTARIO: FOREST FIRE INFO MAP. 2025. Disponível em: <https://www.lioapplications.lrc.gov.on.ca/ForestFireInformationMap/index.html?viewer=FFIM.FFIM>. Acesso em: 25 mai. 2025.

PEDREGOSA, F.; VAROQUAUX, G.; GRAMFORT, A.; MICHEL, V.; THIRION, B.; GRISEL, O.; BLONDEL, M.; PRETTENHOFER, P.; WEISS, R.; DUBOURG, V.; VANDERPLAS, J.; PASSOS, A.; COURNAPEAU, D.; BRUCHER, M.; PERROT, M.; DUCHESNAY, E. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, v. 12, p. 2825–2830, 2011. Acesso em: 12 abr. 2025. Disponível em: <https://scikit-learn.org>.

PEIXOTO, R. *Cerrado: fogo em savanas sobe 221% em agosto; veja ranking de municípios que mais queimaram*. 2024. Acesso em: 6 jun. 2025. Disponível em: <https://g1.globo.com/meio-ambiente/noticia/2024/09/19/cerrado-fogo-em-savanas-sobe-221percent-em-agosto-veja-ranking-de-municipios-que-mais-queimara ghtml>.

PELLEGRINI, A. F. A.; AHLSTRöM, A.; HOBBIE, S. E.; REICH, P. B.; NIERADZIK, L. P.; STAVER, A. C.; SCHARENBROCH, B. C.; JUMPPONEN, A.; ANDEREGG, W. R. L.; RANDERSON, J. T. et al. Fire frequency drives decadal changes in soil carbon and nitrogen and ecosystem productivity. *Nature*, Nature Publishing Group UK London, v. 553, n. 7687, p. 194–198, 2018.

PRIZIBISCZKI, C. *Mesmo em época de chuva, Cerrado tem queimadas acima da média em janeiro*. 2022. Acessado em 6 jun. 2025. Disponível em: <https://oeco.org.br/salada-verde/mesmo-em-epoca-de-chuva-cerrado-tem-queimadas-acima-da-media-em-janeiro/>.

PROGRAMA QUEIMADAS DO INPE. 2025. Disponível em: <https://terrabrasilis.dpi.inpe.br/queimadas/portal/>. Acesso em: 12 mar. 2025.

Python Software Foundation. *Python: programming language*. 2025. <https://www.python.org>. Acesso em: 23 mar. 2025.

RIS-ALA, R. *Fundamentos de Aprendizagem por Reforço*. Rafael Ris-Ala, 2023. ISBN 9786500604368. Disponível em: <https://books.google.com.br/books?id=IKmtEAAAQBAJ>.

RUBí, J. N. S.; GONDIM, P. R. L. A performance comparison of machine learning models for wildfire occurrence risk prediction in the brazilian federal district region. *Environment Systems and Decisions*, v. 44, n. 2, p. 351–368, June 2024. Disponível em: <https://ideas.repec.org/a/spr/envsyd/v44y2024i2d10.1007_s10669-023-09921-2.html>.

SAMUEL, A. L. Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development*, v. 3, n. 3, p. 71–105, 1959.

SCIPY. *SciPy v1.11.3 documentation*. 2023. Acesso em 17 mai. 2025. Disponível em: <https://docs.scipy.org/doc/scipy/>.

SETZER, A. W.; SISMANOGLU, R. A.; SANTOS, J. G. M. dos. Método do cálculo do risco de fogo do programa do inpe - versão 11, junho/2019. *INPE*, 2019. Disponível em: <http://urlib.net/8JMKD3MGP3W34R/3UEDKUB>.

Spyder IDE. *Scientific Python Development Environment*. 2025. <https://www.spyder-ide.org>. Acesso em: 23 mar. 2025.

WASSERMAN, T. N.; MUELLER, S. E. Climate influences on future fire severity: a synthesis of climate-fire interactions and impacts on fire regimes, high-severity fire, and forests in the western united states. *Fire Ecology*, v. 19, n. 1, p. 43, Jul 2023. ISSN 1933-9747. Disponível em: <https://doi.org/10.1186/s42408-023-00200-8>.

WITTEN, I.; FRANK, E. *Data Mining: Practical Machine Learning Tools and Techniques, Second Edition*. Morgan Kaufmann, 2005. (The Morgan Kaufmann Series in Data Management Systems). ISBN 9780080477022. Disponível em: <https://books.google.com.br/books?id=QTnOcZJzlUoC>.

WU, Z.; WANG, B.; LI, M.; TIAN, Y.; QUAN, Y.; LIU, J. Simulation of forest fire spread based on artificial intelligence. *Ecological Indicators*, v. 136, p. 108653, 2022. ISSN 1470-160X. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1470160X22001248>.