Universidade Federal de São Carlos
Centro de Ciências Exatas e de Tecnologia
Departamento de Computação
Bacharelado em Engenharia de Computação

Rômulo Alves da Silva

Previsão de Séries Temporais de Queimadas no Pantanal com uso de Aprendizado de Máquina

São Carlos
2026

Previsão de Séries Temporais de Queimadas no Pantanal com uso de Aprendizado de Máquina

Trabalho de Conclusão de Curso (TCC) apresentado ao curso de Engenharia de Computação, do Centro de Ciências Exatas e de Tecnologia da Universidade Federal de São Carlos , como parte dos requisitos para a obtenção do título de Bacharel em Engenharia de Computação.

Orientador: Marcio Merino Fernandes
Coorientador: Alexandre Levada

São Carlos
2026

Este trabalho é dedicado a todos os estudantes indígenas e pretos da Federal, que ele sirva de inspiração para afirmar que todos nós somos capazes de conquistar aquilo que sonhamos

Agradecimentos

Primeiramente agradeço aos meus pais por me darem todo o suporte ao longo dessa caminhada, que foi árdua e difícil, sem eles eu nunca teria chegado até aqui.

Agradeço também aos amigos que fiz ao longo dessa jornada, sem eles eu não teria ido muito longe nesse curso. Também agradeço ao professor Merino e Levada por todo apoio no desenvolvimento desse trabalho.

Por fim agradeço ao meu mentor e grande amigo professor João Gonçalves, que infelizmente não está mais entre nós. Ele me inspirou a seguir na carreira de engenheiro de computação, e me fez despertar o grande interesse na área da computação e exatas, também sem ele, eu não estaria aqui.

"Isso segue sem parar, sem parar,
o Céu e o Inferno."
(Black Sabbath)

Resumo

O Pantanal brasileiro possui um bioma que é responsável por uma enorme diversidade de fauna e flora. No entanto, as queimadas que afetam esse bioma têm sido um grande problema nos últimos anos, gerando impactos que, futuramente, poderão ser irreversíveis. Desse modo, métodos de previsão de incêndios mostram-se essenciais e de extrema importância para a prevenção e o combate aos incêndios em todo este território. Contudo, os principais índices de risco de incêndio utilizados atualmente apresentam limitações, como: (1) não se ajustarem às características de cada bioma; (2) estarem restritos a variáveis climáticas específicas; e (3) não conseguirem prever o risco de incêndio florestal para um horizonte de dias definido. Este último aspecto, em especial, é de suma importância, pois o seu aprimoramento possibilita o planejamento e a ação coordenada das autoridades ambientais com a devida antecedência.

Como forma de investigar essa limitação, este estudo desenvolveu uma abordagem comparativa entre dois modelos computacionais baseados em Aprendizado de Máquina (XGBoost e LSTM), avaliando qual deles apresentou resultados mais próximos da realidade observada.

Os resultados obtidos indicaram que o modelo XGBoost apresentou desempenho superior na previsão dos focos de incêndio, alcançando menores valores de erro médio absoluto (MAE) e raiz do erro quadrático médio (RMSE), além de maior capacidade de explicação da variabilidade observada na série temporal. O modelo LSTM, por sua vez, demonstrou capacidade de capturar a sazonalidade e o comportamento médio da série, porém apresentou limitações na reprodução da magnitude de eventos extremos, especialmente em função da modelagem univariada e do volume efetivo de dados disponíveis para treinamento. De modo geral, os achados evidenciam que abordagens baseadas em técnicas de boosting associadas à engenharia de atributos temporais mostraram-se mais adequadas para o conjunto de dados analisado, embora ambas as metodologias tenham confirmado a existência de padrões temporais aprendíveis na série histórica do Pantanal brasileiro.

Palavras-chave: Aprendizado de Máquina. XGBoost. LSTM. Pantanal Brasileiro. Queimadas.

Abstract

The Brazilian Pantanal is a biome responsible for an enormous diversity of fauna and flora. However, wildfires affecting this biome have become a major issue in recent years, generating impacts that may become irreversible in the future. In this context, fire prediction methods are essential for prevention and mitigation efforts throughout the region. Nevertheless, the main fire risk indices currently employed present several limitations, such as: (1) lack of adaptation to the specific characteristics of each biome; (2) restriction to specific climatic variables; and (3) inability to forecast fire risk for a defined time horizon. The latter aspect is particularly important, as its improvement enables environmental authorities to plan and coordinate actions in advance.

As a way to investigate this limitation, this study developed a comparative approach between two computational models based on Machine Learning (ML), XGBoost and LSTM, evaluating which of them presented results closer to the observed reality.

The results indicated that the XGBoost model achieved superior predictive performance, presenting lower Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE), as well as greater explanatory capacity of the variability observed in the time series. The LSTM model demonstrated the ability to capture seasonality and the average behavior of the series; however, it showed limitations in reproducing the magnitude of extreme events, particularly due to the univariate modeling approach and the limited effective training sample size. Overall, the findings suggest that boosting-based approaches combined with temporal feature engineering are more suitable for the analyzed dataset, although both methodologies confirmed the presence of learnable temporal patterns in the historical wildfire series of the Brazilian Pantanal.

Keywords: Machine Learning. Gradient Boosting. Long Short-Term Memory. Brazilian Pantanal. Wildfire Prediction..

Lista de ilustrações

Figura 1 – Abordagens de Aprendizado de Máquina. 20
Figura 2 – Diferença entre problemas de classificação: (a) Plano; (b) Hierárquico. 21
Figura 3 – Ilustração de Árvore de Decisão. Adaptado de (MITCHELL, 1997). . . 23
Figura 4 – Ilustração do XGBoost. 24
Figura 5 – Ilustração do Perceptron. 26
Figura 6 – Ilustração do MLP. 27
Figura 7 – Ilustração do RNN. 29
Figura 8 – Ilustração de um LSTM. c_{prev} representa o estado anterior da célula, h_{prev} representa o estado oculto anterior, c_{curr} representa o estado atual da célula e h_{curr} representa o estado oculto atual. Adaptado de (ZHAO et al., 2020). 30
Figura 9 – Comparação do total de focos ativos detectados pelo satélite de referência em cada mês, no período de 1998 até 26/jan. 44
Figura 10 – Histórico de Focos de Incêndio (1998–2025). 52
Figura 11 – Padrões Sazonais e Distribuição Temporal. 56
Figura 12 – Padrões Sazonais e Distribuição Temporal. 57
Figura 13 – Padrões Sazonais e Distribuição Temporal. 58
Figura 14 – Análise dos Resíduos. 59
Figura 15 – Previsões versus Valores Reais. 60
Figura 16 – Dashboard de Previsão de Queimadas. 61
Figura 17 – Previsões versus Valores Reais. 63
Figura 18 – Análise de Resíduos. 64
Figura 19 – LSTM: Previsões versus Valores Reais. 65

Lista de tabelas

Tabela 1 – Ajustes no cálculo do FMA conforme a precipitação diária. Adaptado de (SOARES; BATISTA, 2007). 32
Tabela 2 – Classificação do risco de incêndio florestal com base na FMA, Adaptado de (SOARES; BATISTA, 2007) 32
Tabela 3 – Classificação do risco de incêndio florestal com base na FMA^{+} 33
Tabela 4 – Ajustes no cálculo do índice Telicyn de acordo com a precipitação diária 34
Tabela 5 – Classes de Risco de Incêndio Florestal de acordo com o índice Telicyn . 34
Tabela 6 – Classes de Risco de Incêndio Florestal de acordo com o índice Telicyn . 34
Tabela 7 – Classes de risco de incêndio florestal de acordo com o índice Ângström 35
Tabela 8 – Ajustes no cálculo do índice de Nesterov conforme a precipitação . . . 36
Tabela 9 – Classes de risco de incêndio florestal de acordo com o índice de Nesterov.Adaptado de (SOARES; BATISTA, 2007) 36
Tabela 10 – Variáveis climáticas exigidas para cada um dos principais índices de risco de incêndio florestal 37
Tabela 11 – Variáveis climáticas exigidas para cada um dos principais índices de risco de incêndio florestal. 37

Lista de siglas

ADASYN Adaptive Synthetic Sampling

AI Artificial Intelligence

AM Aprendizado de Máquina

ANN Artificial Neural Network

ARIMA Autoregressive Integrated Moving Average

CatBoost Categorical Boosting

CNN Convolutional Neural Network

DL Deep Learning

ENN Edited Nearest Neighbours

FMA Fórmula de Monte Alegre

FMA$^+$ Fórmula de Monte Alegre Modificada

FWI Fire Weather Index

GBDT Gradient Boosting Decision Tree

GRU Gated Recurrent Unit

IA Inteligência Artificial

INPE Instituto Nacional de Pesquisas Espaciais

LightGBM Light Gradient Boosting Machine

LSTM Long Short-Term Memory

MAE Mean Absolute Error (Erro Médio Absoluto)

MAPE Mean Absolute Percentage Error

ML Machine Learning

MLP Multi-Layer Perceptron

MSE Mean Squared Error

NDVI Normalized Difference Vegetation Index

RF Random Forest

RMSE Root Mean Square Error

RNN Recurrent Neural Network

R^2 Coeficiente de Determinação

SARIMA Seasonal Autoregressive Integrated Moving Average

SARIPAN Sistema de Avaliação de Risco de Incêndio para o Pantanal

SMOTE Synthetic Minority Over-sampling Technique

XGBoost Extreme Gradient Boosting

Sumário

1 INTRODUÇÃO 13
1.1 Contextualização e Motivação 13
1.2 Motivação e Justificativa 14
1.3 Objetivos do Trabalho 16
1.3.1 Objetivo Geral 16
1.3.2 Objetivos Específicos 16
1.4 Organização do Texto 17

2 FUNDAMENTAÇÃO TEÓRICA 18
2.1 Pré-processamento de Dados 18
2.1.1 Normalização (Scaling) 18
2.2 Aprendizado de Máquina (AM) 19
2.2.1 Paradigma Simbólico 21
2.2.2 Paradigma Conexcionista 25
2.3 Previsão de Séries Temporais (TS) 27
2.3.1 Rede Neural Recorrente (RNN) 28
2.3.2 Índices de Riscos Florestais 31
2.3.3 Fórmula de Monte Alegre (FMA) e Fórmula de Monte Alegre Modificada (FMA^{+}) 31
2.3.4 Telicyn 33
2.3.5 Angström 34
2.3.6 Nesterov 35
2.3.7 Comparação dos Índices 37
2.3.8 Medidas Preventivas 38

3 METODOLOGIA E DESENVOLVIMENTO DO PROJETO . 41
3.1 Justificativa Formal da Escolha dos Modelos 41
3.1.1 Justificativa da Escolha do XGBoost 41
3.1.2 Justificativa da Escolha da LSTM 42
3.1.3 Fundamentação Comparativa 43
3.2 Variáveis de Observações 43
3.3 Estratégia de Divisão Temporal e Validação de Dados 44
3.3.1 Divisão para o XGBoost 45
3.3.2 Divisão para LSTM 45
3.3.3 Garantia de Integridade Temporal 45
3.4 Implementação XGBoost 46

3.4.1 Organização do Pipeline de Processamento 46
3.4.2 Formulação Conceitual do Modelo XGBoost no Contexto Temporal . . . 47
3.4.3 Estratégia de Previsão Futura (Multi-step Forecast) 47
3.5 Implementação LSTM 48
3.5.1 Organização do Pipeline de Processamento 49
3.5.2 Formulação Conceitual do Modelo LSTM no Contexto Temporal 50
3.5.3 Estratégia de Previsão Futura (Multi-step Forecast) 50

4 EXPERIMENTOS E ANÁLISE DE RESULTADOS 52
4.1 Análise Exploratória da Série Temporal 52
4.1.1 Histórico de Focos de Incêndio 52
4.1.2 Padrões Sazonais e Distribuição Temporal 53
4.2 Engenharia de Atributos e Importância de Variáveis 56
4.2.1 Importância das Features 56
4.3 Avaliação Preditiva do Modelo 57
4.3.1 Previsões versus Valores Reais 57
4.3.2 Análise dos Resíduos 58
4.4 Avaliação Temporal das Previsões 59
4.4.1 Previsões ao Longo do Tempo 59
4.5 Dashboard Integrado de Resultados 60
4.6 Discussão Geral dos Resultados 61
4.6.1 Avaliação Preditiva do modelo LSTM 62
4.6.2 Análise dos Resíduos 63
4.6.3 Avaliação Temporal das previsões 64
4.6.4 Discussão Geral dos Resultados 65
4.6.5 Comparação entre os modelos LSTM e XGBoost 66

5 CONCLUSÃO 68

REFERÊNCIAS 70

1 Introdução

1.1 Contextualização e Motivação

O Pantanal constitui um dos maiores ambientes úmidos do planeta, abrangendo cerca de 150.355 km², dos quais a maior parte deste território se encontra no Brasil, estendendo-se também para a Bolívia e pelo Paraguai. Sua relevância ecológica é amplamente reconhecida, sendo classificado pela UNESCO (Organização das Nações Unidas para a Educação, a Ciência e a Cultura) como Reserva da Biosfera. Embora ainda preserve aproximadamente 80% de sua vegetação nativa (ALHO et al., 2019), os ecossistemas pantaneiros sofrem constantes pressões decorrentes de atividades humanas, somadas às intensas variações sazonais de seu regime climático.

Entre as principais ameaças ao Pantanal destacam-se os incêndios florestais, fenômeno fortemente relacionado às condições climáticas. Durante os meses de junho a setembro, a região enfrenta um período de escassez de chuvas, o que favorece o acúmulo de material vegetal seco. Esse acúmulo, por sua vez, atua como combustível, aumentando significativamente o risco de ignição e propagação do fogo nos ecossistemas pantaneiros.

A utilização de dados meteorológicos possibilita a classificação do risco de ocorrência de incêndios florestais, constituindo uma ferramenta importante para ações de prevenção e controle. Em (SORIANO; DANIEL; SANTOS, 2015), foram avaliadas as eficiências de cinco diferentes índices aplicados ao Pantanal brasileiro: Fórmula de Monte Alegre (FMA), Fórmula de Monte Alegre Modificada (FMA+), Telicyn, Ångström e Nesterov. De modo semelhante, o estudo de (TORRES; RIBEIRO, 2008) compara os índices FMA, Telicyn, Nesterov, Índice de Precipitação-Evaporação (P-EVAP) e Índice de Evaporação Cumulativa por Precipitação (EVEAP/P) com o intuito de prever incêndios florestais no município de Juiz de Fora (MG). Entre os diversos indicadores, a FMA destaca-se como um dos mais utilizados no Brasil, tendo sido originalmente desenvolvida a partir de estudos na região de Araucária. No Brasil o índice FMA foi desenvolvido por (SOARES, 1972) e caracteriza-se por ser cumulativo, considerando apenas duas variáveis em sua fórmula: umidade relativa e precipitação. Esse índice tem sido amplamente empregado em diferentes regiões do país para a classificação do risco de incêndios florestais (TETTO et al., 2010; NUNES et al., 2010; ALVARES et al., 2014), incluindo o Pantanal (SORIANO; DANIEL; SANTOS, 2015; ONIGEMO, 2007).

Além disso, destaca-se o sistema SARIPAN (NARCISO; SORIANO, 2019), atualmente utilizado por órgãos ambientais no combate aos incêndios florestais no Pantanal brasileiro. Esse sistema integra os índices FMA, FMA+, Nesterov, Telicyn e Ångström para identificação do risco de ocorrência de fogo.

Apesar da sua relevância, a aplicação desses índices apresenta algumas limitações, como:

1. Ausência de ajustes às particularidades de cada bioma;
2. Restrição a um conjunto limitado de variáveis climáticas;
3. Incapacidade de prever os riscos de incêndios florestais em um horizonte temporal definido.

1.2 Motivação e Justificativa

Na classificação do risco de incêndios florestais, uma alternativa promissora consiste na utilização de modelos fundamentados em Aprendizado de Máquina (AM). Diferentemente dos índices tradicionais, esses modelos apresentam vantagens de incorporar de forma mais precisa as particularidades da região em análise, além de não imporem restrições quanto ao número ou ao tipo de variáveis utilizadas para representar os fatores ambientais e climáticos.

Pesquisas recentes evidenciam que o uso de algoritmos baseados em Redes Neurais Artificiais (RNA) tem apresentado resultados promissores na detecção de incêndios florestais (LUO et al., 2018; AL-ZEDA et al., 2021; YANG; LUPASCU; MEEL, 2021; GAO; LIN; LIN; HU, 2023), incluindo aplicações específicas no Pantanal brasileiro (VIGANÓ et al., 2017). Nesse contexto, (Rakshit et al., 2021) avaliaram diferentes algoritmos de Aprendizado de Máquina na classificação de áreas de acordo com o grau de suscetibilidade a incêndios - altamente propensas, moderadamente propensas, pouco propensas ou não propensas. Os classificadores testados incluíram Árvores de Decisão, K-Vizinhos Mais Próximos (KNN), Máquinas de Vetores de Suporte (SVM) e Naive Bayes, sendo as Árvores de Decisão aquelas que obtiveram o melhor desempenho.

Um estudo recente conduzido por (RUBÍ; CARVALHO; GONDIM, 2023) analisou a aplicação de diferentes modelos de Aprendizado de Máquina na previsão tanto na propagação quanto do comportamento dos incêndios na região do Distrito Federal, que faz parte do bioma pertencente ao Cerrado. O conjunto de variáveis explicativas utilizado incluiu fatores climáticos, dados de sensoriamento remoto, além de informações topográficas, hidrográficas e antrópicas. Em relação à previsão da propagação do fogo, os resultados indicaram que o modelo AdaBoost apresentou desempenho superior aos demais algoritmos avaliados - como Random Forest, Redes Neurais Artificiais (RNA) e Máquinas de Vetores de Suporte (SVM) - alcançando 91% de acurácia.

Entretanto, observa-se que os estudos mencionados ainda apresentam limitações na estimativa do risco de incêndios florestais em horizontes temporais de curto prazo, especialmente em escalas diárias. Essa restrição reduz seu potencial de aplicação no planejamento

estratégico e operacional, uma vez que dificulta a geração de informações antecipadas, com alguns dias ou poucas semanas de antecedência, que possam subsidiar a organização de ações preventivas e o apoio à tomada de decisão pelos órgãos competentes.

Nesse contexto, a aplicação de técnicas de previsão de Séries Temporais (ST) mostra-se viável, pois os dados de focos de incêndio estão organizados cronologicamente e apresentam padrões temporais, como sazonalidade e tendência, que podem ser explorados para estimar comportamentos futuros. Uma ST é composta por dados registrados ao longo do tempo referentes à observação de um determinado fenômeno, como variáveis climáticas. Para realização de previsões a partir dessas séries, podem ser empregados métodos estatísticos baseados nas autocorrelações presentes nos dados, como modelos Autorregressivos Integrado de Médias Móveis (ARIMA) e Autorregressivo Integrado de Médias Móveis Sazonais (SARIMA) (DIMRI; AHMAD; SHARIF, 2020; MURAT et al., 2018; MUKADI; GONZÁLEZ-GARCÍA, 2021).

Além desses métodos tradicionais, é possível recorrer também a modelos fundamentados em Aprendizado de Máquina, como as redes de Memória de Longo Prazo (LSTM) (ABBES; MAGAGI; GOITA, 2019). Nesse sentido, (LIN et al., 2023) aplicaram LSTM para a previsão de incêndios florestais na região de Chogli, China; entretanto, o estudo limitou-se a prever apenas um dia à frente, não abrangendo horizontes temporais mais longos.

Desta forma, evidencia-se que modelos de Aprendizado de Máquina podem ser explorados tanto para a previsão de variáveis climáticas quanto para a classificação de riscos de incêndios florestais. Contudo, diante da ampla variedade de algoritmos e abordagens disponíveis, a busca pelo ajuste ideal de hiperparâmetros configura-se como um desafio relevante. Nesse cenário, os Algoritmos Genéricos (AG) surgem como uma alternativa eficiente, com potencial para alcançar resultados superiores.

Nesse sentido (GANAPATHY, 2020) propôs o uso de AG como método automático de ajustes para redes neurais, obtendo desempenho superior a uma busca aleatória de hiperparâmetros em uma tarefa de tradução automática do japonês para o inglês. De forma complementar, (ALIBRAHIM; LUDWIG, 2021) comparam o uso de AG com Grid Search e Otimização Bayesiana em um problema de previsão de transações de clientes, observando novamente melhores resultados com o AG - que alcançou 0,826 de Área Sob a Curva ROC (AUC), frente a 0,792 de Grid Search e 0,789 de Otimização Bayesiana.

Modelos baseados em AG também vêm sendo explorados em aplicações ambientais e climáticas. (AMOL, 2020), por exemplo, apresentou um AG para prever a programação de incêndios florestais, considerando quatro parâmetros; fator seca, temperatura, umidade relativa e vento. O modelo identificou intervalos críticos de valores desses parâmetros que favorecem a ignição do fogo, mostrando-se preciso na previsão da categoria de propagação da área afetada. Já o estudo de (MATOS et al., 2022) aplicou AG ao problema de escalonamento de recursos no combate a incêndios florestais, buscando determinar a melhor

sequência de ações a serem executadas por equipes de enfrentamento. Utilizando dados coletados em Braga, Portugal, os resultados confirmaram a viabilidade e relevância da abordagem.

Ainda assim, até onde se tem conhecimento, existem poucos estudos que apliquem Algoritmos Genéricos especificamente à seleção de hiperparâmetros voltada à previsão de variáveis climáticas e à classificação do risco de incêndios florestais.

1.3 Objetivos do Trabalho

1.3.1 Objetivo Geral

O objetivo deste Trabalho de Conclusão de Curso pode ser apresentado da seguinte forma:

Diante desse contexto, a presente pesquisa busca responder à seguinte questão: em que medida séries históricas de focos de incêndio, associadas às variáveis climáticas da região, podem ser utilizadas para prever o risco de ocorrência de incêndios no Pantanal brasileiro com antecedência temporal útil ao apoio à tomada de decisão?

O objetivo geral deste trabalho é desenvolver um sistema de comparação entre dois métodos computacionais, um baseado em árvores de decisão e o outro em deep learning, capazes de:

1. Prever os focos de incêndio no Pantanal brasileiro para horizontes temporais específicos, utilizando séries históricas e variáveis associadas ao fenômeno.

2. Comparar os modelos gerados pelos dois métodos computacionais usando gráficos de representação e medir a sua efetividade ao longo do tempo, e medir sua taxa de acerto em relação aos fatos do mundo real.

1.3.2 Objetivos Específicos

1. Organizar e preparar a base histórica de dados de focos de incêndio utilizada no estudo, de modo a viabilizar sua aplicação em modelos preditivos;

2. Construir uma abordagem preditiva com o modelo XGBoost, a partir da reformulação da série temporal em uma base supervisionada com atributos derivados;

3. Desenvolver uma abordagem preditiva com rede neural LSTM, considerando sua capacidade de modelar dependências temporais em séries históricas;

4. Comparar o desempenho dos modelos XGBoost e LSTM por meio de métricas de avaliação e da análise de sua capacidade de representar o comportamento temporal dos focos de incêndio;

5. Analisar os resultados obtidos, identificando potencialidades e limitações de cada abordagem para a previsão de focos de incêndio no Pantanal brasileiro.

1.4 Organização do Texto

O restante deste documento está organizado da seguinte forma:

- Capítulo 2 - Fundamentação Teórica
  - 2.1 - Pré-processamento de Dados: apresentar os principais métodos de pré-processamento empregados, com ênfase em normalização. São descritos os procedimentos implementados no âmbito do projeto.
  - 2.2 - Aprendizado de Máquina: introduz conceitos fundamentais de Aprendizado de Máquina aplicados à classificação de dados, paradigmas que são utilizados na literatura e nesse estudo, bem como algoritmos implementados para conclusão desse estudo.
  - 2.3 - Previsão de Séries Temporais: aborda os conceitos de séries temporais e discute os principais algoritmos utilizados na literatura para esse tipo de previsão, destacando aqueles implementados neste projeto.
  - 2.4 - Índices de Risco de Incêndio Florestal: apresenta os principais índices de risco descritos na literatura, suas formas de aplicação e limitações identificadas.
- Capítulo 3 - Materiais e Métodos: detalha os dados empregados e os métodos utilizados para a construção do sistema de comparação proposto.
- Capítulo 4 - Experimentos e Análise de Resultados.
  - 4.1 - Sistema: descreve os pipelines de treinamento e previsão do sistema, destacando as diferenças entre os dois modelos utilizados neste método de comparação.
  - 4.2 - Experimentos e Resultados: reúne os experimentos conduzidos com o sistema de comparação, bem como os resultados obtidos.
- Capítulo 5 - Conclusão: apresenta as principais conclusões, discute resultados obtidos e propõe perspectivas para trabalhos futuros.

2 Fundamentação Teórica

Este capítulo apresenta os fundamentos teóricos que sustentam a modelagem adotada neste estudo. Inicialmente, é discutida a etapa de pré-processamento de dados, com ênfase na normalização das variáveis, procedimento importante para a preparação da base utilizada nos experimentos. Em seguida, são apresentados os conceitos centrais de Aprendizado de Máquina, com foco no aprendizado supervisionado e em sua aplicação à previsão de séries temporais. Na sequência, são abordadas as duas estratégias de modelagem investigadas neste trabalho: modelos baseados em árvores de decisão, representados pelo XGBoost, e redes neurais recorrentes, com destaque para a arquitetura Long Short-Term Memory (LSTM). Por fim, são discutidos os principais índices de risco de incêndio florestal descritos na literatura, ressaltando suas contribuições e limitações no contexto do Pantanal brasileiro.

2.1 Pré-processamento de Dados

Nos diversos Problemas de Ciência de Dados (CD), uma das etapas iniciais e mais relevantes do processo é o pré-processamento de dados. Essa fase consiste na aplicação de técnicas que permitem adaptar e transformar dados, tornando-os adequados para alimentar modelos de Aprendizado de Máquina (AM), conforme será aprofundado na seção 2.2.

Segundo (XIANG-WEI; YIAN-FANG, 2012), o pré-processamento de dados constitui uma etapa crítica em diferentes aplicações de CD, uma vez que um conjunto de dados devidamente tratado pode não apenas aumentar a acurácia dos modelos de AM, mas também melhorar significativamente sua eficiência computacional.

De modo geral, o pré-processamento envolve procedimentos como limpeza, integração, transformação e redução de dados, entre outros. No presente estudo, destacam-se duas técnicas principais de pré-processamento: normalização (scaling) e amostragem (sampling), que serão detalhadas nas subseções seguintes.

2.1.1 Normalização (Scaling)

A normalização de dados - ou mais especificamente, de features - refere-se ao método de pré-processamento que atualiza a faixa de valores das variáveis independentes do conjunto de dados. É comumente confundida com o termo "normalização", porém este é apenas um dos diferentes tipos de scaling.

Alguns algoritmos de Aprendizado de Máquina (AM) são muito sensíveis à faixa de dados, especialmente quando existem variáveis com amplitudes muito diferentes entre si.

Por isso, os métodos de scaling desempenham um papel importante no treinamento dos modelos de AM e nos ajustes de dados.

O estudo de (AMBARWARI;ADRIAN;HERDIYENI,2020) demonstrou que métodos de scaling fornecem melhorias significativas no desempenho de algoritmos de AM, como KNN, Naive Bayes, RNA, SVM. Os resultados revelam que o SVM com normalização superou o desempenho de outros algoritmos. Da mesma forma, (AHSAN et al., 2021) investigaram os efeitos de diferentes métodos de scaling no desempenho de modelos de AM. Seus resultados mostram que a performance dos algoritmos varia de acordo com o método de normalização utilizado.

Não foi possível identificar um único método de scaling que pudesse ser classificado como o melhor entre todos.

Conforme será detalhado posteriormente no Capítulo 4 na aba de Sistemas, neste projeto foram investigados os seguintes métodos de normalização:

1. Min-Max: Escala cada feature de forma que seus valores fiquem entre 0 e 1;
2. Max-Abs: Escala cada feature de forma que o valor absoluto máximo de cada variável seja igual a 1;
3. Robust: Escala as variáveis removendo a mediana e ajustando-as de acordo com o intervalo interquartil;
4. Standard: Escala as variáveis removendo a média e ajustando para variância unitária;

2.2 Aprendizado de Máquina

O Aprendizado de Máquina (AM), subárea da Inteligência Artificial (IA), dedica-se ao estudo e desenvolvimento de algoritmos capazes de permitir que sistemas computacionais melhorem automaticamente seu desempenho a partir da experiência (MITCHELL, 1997). Em essência, o AM busca substituir ou complementar a programação tradicional, típica da computação convencional, pela aprendizagem baseada em dados (RUSSELL; NORVIG, 2010; ALPAYDIN, 2014). Na computação tradicional, a resolução de um problema específico envolve duas etapas principais: (1) A Aquisição de conhecimento sobre o domínio do problema; (2) A elaboração de algoritmos que incorporem regras de negócio, funções e/ou modelos matemáticos voltados à solução desejada. Esse processo, embora estruturado, apresenta limitações significativas, como baixa capacidade de adaptação a mudanças no domínio do problema e o elevado custo de tempo e recursos humanos. Com o Aprendizado de Máquina, por outro lado, o processo é reorganizado em etapas mais flexíveis baseadas em dados: (1) Coleta e limpeza dos dados, que descrevem o domínio

do problema; (2) Treinamento de algoritmos, como classificadores ou modelos de regressão para identificar padrões e realizar previsões. Dessa forma, a tarefa de aprendizado é delegada à máquina, que passa a construir o modelo de forma automática com base nos dados fornecidos. As abordagens de AM variam conforme o tipo e a estrutura dos dados disponíveis. A figura 1 apresenta uma visão geral das principais abordagens.

Entre essas abordagens, destaca-se o Aprendizado Supervisionado, que utiliza conjunto de dados rotulados. Nesse tipo de aprendizado, cada exemplo é associado a um rótulo pertencente a um conjunto de classes $y \subset Y$, sendo Y o domínio completo das classes (no caso da classificação) ou uma saída numérica contínua (no caso de regressão).

No presente estudo, o problema é tratado no contexto de Aprendizado Supervisionado, com foco em uma tarefa de previsão de séries temporais formulada em termos de regressão supervisionada. Em vez de atribuir cada exemplo a classes discretas de ocorrência ou não ocorrência de incêndio, busca-se estimar o comportamento futuro da variável de interesse a partir de seus valores históricos e de atributos temporais derivados. Assim, a modelagem proposta está voltada à previsão dinâmica temporal dos focos de incêndio no Pantanal brasileiro, em consonância com os objetivos e resultados apresentados neste trabalho.

Dessa forma, busca-se encontrar uma função capaz de mapear o conjunto de atributos de entrada para uma saída numérica, correspondente à estimativa do valor futuro da série. Esse modelo é construído a partir de um conjunto de treinamento no qual os valores de saída são conhecidos para os exemplos observados. O modelo pode ser representado de diferentes formas, como árvores de decisão, métodos baseados em ensemble ou redes neurais. No contexto deste trabalho, o XGBoost é empregado a partir da reformulação da série temporal em uma base supervisionada com atributos derivados, enquanto a LSTM é utilizada para modelar diretamente as dependências sequenciais dos dados ao longo do tempo (ambos os modelos são explicados mais à frente nesse trabalho).

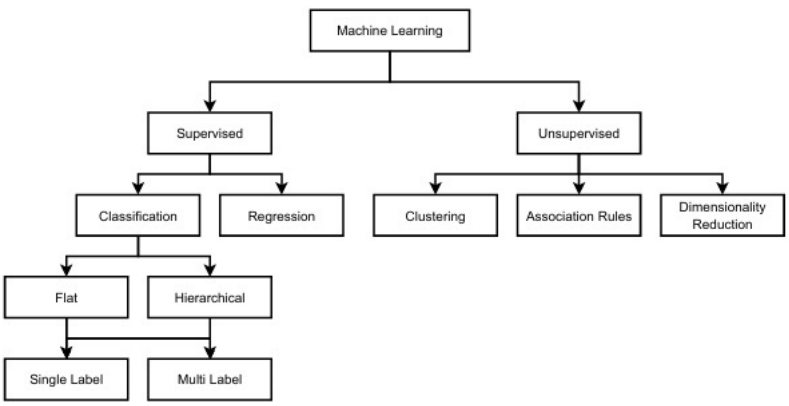

Figura 1 – Abordagens de Aprendizado de Máquina.

Embora o foco deste trabalho esteja em regressão supervisionada aplicada à previsão temporal, apresenta-se, a seguir, uma breve contextualização sobre problemas de classifi-

cação, por sua relevância na literatura de Aprendizado de Máquina e para compreensão geral das abordagens supervisionadas.

No caso em que o conjunto de classes possui apenas duas possibilidades, o problema é classificado como binário. Quando o número de classes é maior que dois, tem-se um problema de múltiplas classes. Já quando um mesmo objeto pode ser associado simultaneamente a mais de uma classe, o problema é denominado multirrótulo.

Na literatura de Aprendizado de Máquina, os problemas de classificação tradicionais são, em geral, resolvidos por meio de métodos de classificação plana (flat ou não hierárquica). Nessa abordagem, as classes são tratadas de forma independente, desconsiderando eventuais relações hierárquicas entre elas.

Em contrapartida, em problemas de classificação hierárquica, as classes são organizadas em uma estrutura taxonômica, podendo ser divididas em subclasses e agrupadas em superclasses. Nesses casos, os classificadores consideram as relações entre os níveis hierárquicos para realizar a predição. Assim, o classificador f deve respeitar as restrições impostas pela taxonomia, o que significa que, ao prever uma classe específica, todas as suas superclasses também devem ser incluídas na predição. A figura 2 ilustra a diferença entre os problemas de classificação plana e classificação hierárquica.

Nas seções seguintes deste capítulo, são apresentados alguns dos principais paradigmas de Aprendizado de Máquina, com ênfase em métodos supervisionados amplamente utilizados tanto em tarefas de classificação quanto de regressão. Embora a classificação seja discutida como parte da fundamentação teórica geral, o foco deste estudo recai sobre a previsão de séries temporais, formulada como um problema de regressão supervisionada.

2.2.1 Paradigma Simbólico

O Paradigma Simbólico fundamenta-se na representação de conceitos do mundo real por meio de símbolos, que são manipulados a partir de regras explícitas. Essas regras permitem tanto a inferência quanto a generalização do conhecimento, sendo geralmente passíveis de interpretação em linguagem natural, o que confere transparência ao processo de aprendizado.

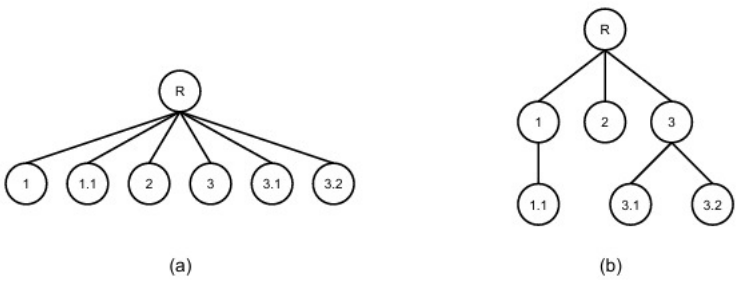

Figura 2 – Diferença entre problemas de classificação: (a) Plano; (b) Hierárquico.

Entre os algoritmos que se enquadram nesse paradigma, destacam-se as Árvores de Decisão, que estruturam o conhecimento em forma de regras condicionais, e seus modelos derivados mais complexos, como as Florestas Aleatórias (Random Forest), que combinam múltiplas árvores para melhorar a precisão e reduzir a variância das previsões.

2.2.1.1 Árvores de Decisão

As Árvores de Decisão constituem uma categoria de algoritmos de Aprendizado de Máquina supervisionado amplamente empregada em diferentes domínios, destacando-se pela simplicidade, eficiência e facilidade de interpretação. Nesses modelos, a estrutura é representada na forma de uma árvore hierárquica, na qual cada nó interno corresponde a um teste aplicado sobre um atributo, e cada nó-folha representa uma classe de saída.

O processo de classificação ocorre de maneira sequencial: para cada objeto a ser classificado, percorre-se a árvore desde o nó raiz até um nó-folha, aplicando, em cada nível, o teste correspondente ao atributo definido naquele nó. A classe associada ao nó-folha alcançado representa o resultado final da classificação para o objeto analisado.

A figura 3 ilustra um exemplo de Árvore de Decisão construída a partir de dados climáticos, considerando atributos de umidade, temperatura e velocidade do vento. A variável de saída é binária, indicando se ocorrerá ou não chuva ("Sim"ou "Não").

Por exemplo, considere um objeto com os seguintes valores de atributos: Umidade = Alta, Vento = Fraco, Temperatura = Alta e Céu = Nublado. Ao percorrer a Árvore de Decisão conforme esses valores, obtém-se o resultado "Não"indicando a ausência de chuva naquele dia.

Diversos algoritmos de construção de Árvores de Decisão foram propostos na literatura, diferenciando-se pelas heurísticas e métodos de particionamento utilizados. Entre os mais conhecidos, destacam-se o ID3 (QUINLAN, 1986) e seu sucessor, o C4.5 (QUINLAN, 1993).

Em geral, a construção de uma Árvore de Decisão segue um procedimento top-down (de cima para baixo), iniciando-se pela definição do nó raiz e avançando recursivamente para os nós subsequentes. A primeira questão a ser respondida nesse processo é: "Qual atributo deve ser testado na raiz da árvore?"(MITCHELL, 1997). Para isso, aplicam-se métricas heurísticas a cada atributo, selecionando aquele que melhor separa os exemplos pertencentes a diferentes classes. O objetivo é maximizar a pureza dos subconjuntos resultantes, garantindo maior poder discriminativo na classificação dos nós dessa Árvore.

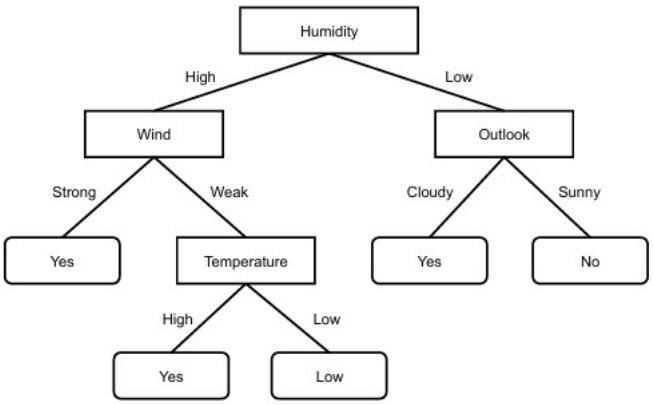

Figura 3 – Ilustração de Árvore de Decisão. Adaptado de (MITCHELL, 1997).

Durante a classificação da Árvore de Decisão, os objetos que satisfazem a condição estabelecida em determinado nó são direcionados para um dos ramos, enquanto aqueles que não satisfazem seguem para o outro. Esse processo é, em geral, baseado em divisões binárias.

Para divisões não binárias, o procedimento é análogo, diferenciando-se apenas pelo fato de que cada nó pode gerar mais de dois ramos filhos. O processo de particionamento é então repetido de forma recursiva, até que seja alcançado um critério de parada pré-definido, resultando em um nó-folha. A classe final atribuída a esse nó-folha corresponde, normalmente, à classe majoritária entre os objetos que chegaram até ele.

Uma das principais vantagens das Árvores de Decisão em tarefas de classificação é sua alta interpretabilidade. A partir da visualização a estrutura da árvore, torna-se possível identificar quais atributos exercem maior influência sobre a determinação das classes, além de compreender as decisões tomadas em cada nível do modelo.

O custo computacional das Árvores de Decisão depende diretamente das heurísticas e dos parâmetros definidos, como a profundidade máxima da árvore ou o número mínimo de amostras por nó. Ademais, conjuntos de dados com grande quantidade de atributos contínuos tendem a aumentar a complexidade dos cálculos de seleção de atributos e divisão de nós.

Por fim, é importante ressaltar que as Árvores de Decisão são sensíveis a dados desbalanceados, podendo gerar modelos tendenciosos em favor da classe dominante. Nessas situações, recomenda-se o uso de técnicas de balanceamento de dados - como os métodos de pré-processamento discutidos na seção 2.1 - para mitigar o viés do classificador.

2.2.1.2 XGBoost

O XGBoost é um algoritmo do tipo ensemble baseado em Árvores de Decisão (ver Seção 2.2.1.1), que utiliza a técnica de boosting para aprimorar o desempenho do modelo. O método foi proposto por (CHEN; GUESTRIN,2016) e tornou-se amplamente adotado

em competições e aplicações práticas de Aprendizado de Máquina devido à sua eficiência computacional e elevada precisão.

O princípio fundamental do boosting consiste em construir um classificador forte a partir da combinação sequencial de múltiplos classificadores fracos. Em cada iteração, o modelo busca corrigir os erros cometidos pelos classificadores anteriores, de modo que o desempenho global seja progressivamente aprimorado.

Entre os principais algoritmos que utilizam a técnica de boosting, destacam-se o AdaBoost (SCHAPIRE,2013), o LightGBM (KE et al., 2017) (ver Seção 2.2.1.4), o CatBoost (PROKHORENKOVA et al., 2018) (ver Seção 2.2.1.5) e o próprio XGBoost.

O XGBoost emprega uma variação específica denominada Gradient Boosting, cujo objetivo é encontrar uma função f que melhor descreva os dados a determinar os parâmetros ótimos p dessa função. Para isso, o modelo combina diversas funções bases simples, geralmente Árvores de Decisão de pequena profundidade, utilizando como entrada o gradiente do erro em relação às previsões anteriores. Dessa forma, cada nova função é treinada para corrigir os erros residuais do modelo anterior, permitindo que o conjunto final aprenda de maneira mais eficaz as relações subjacentes aos dados.

Em síntese, no XGBoost, os classificadores são ajustados em série, e a cada iteração os erros residuais do classificador precedente são incorporados ao treinamento do próximo. A classificação final é obtida por meio da soma ponderada das saídas de todos os classificadores fracos, resultando em um modelo robusto e de alta capacidade preditiva.

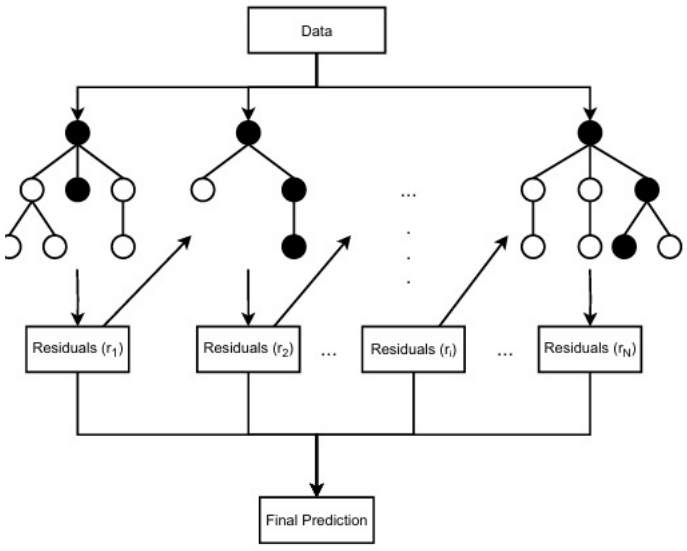

Figura 4 – Ilustração do XGBoost.

No XGBoost, as árvores são construídas de forma paralela, o que garante alta eficiência computacional e contribui para seu desempenho superior em relação a diversos outros algoritmos de Aprendizado de Máquina reportados na literatura. O modelo apresenta resultados consistentes tanto em conjuntos de dados reduzidos quanto em bases de grande

escala, demonstrando elevada capacidade de generalização. No entanto, seu desempenho pode ser comprometido em cenários com dados altamente esparsos ou fortemente desbalanceados.

De maneira análoga ao observado na Floresta Aleatória (ver Seção 2.2.1.2), o uso do XGBoost implica a perda de interpretabilidade característica das Árvores de Decisão individuais, uma vez que o modelo final resulta da combinação de múltiplas estruturas em sequência, dificultando a compreensão direta das regras de decisão aprendidas

2.2.2 Paradigma Conexcionista

O Paradigma Conexcionista tem como principal inspiração a estrutura biológica do cérebro humano, composta por um grande número de unidades interconectadas (neurônios) que se comunicam por meio de sinais elétricos e sinapses.

A partir desse paradigma, foram desenvolvidas as Redes Neurais Artificiais (RNAs), modelos computacionais capazes de aprender padrões complexos por meio da interação entre múltiplas unidades artificiais (neurônios), as quais funcionam de maneira análoga ao processamento de informações realizado pelo sistema nervoso biológico.

2.2.2.1 Perceptron Multicamadas (MLP)

O perceptron é considerado o modelo mais simples de um neurônio artificial, tendo sido proposto pela primeira vez por (ROSENBLATT, 1957) em seu trabalho seminal. Esse modelo constitui a base conceitual das Redes Neurais Artificiais (RNAs) modernas e é ilustrado na Figura 5.

O processo de treinamento básico de um perceptron pode ser descrito pelas seguintes etapas:

1. Inicialização dos pesos:

   Defini-se aleatoriamente um vetor de pesos

   $$\mathbf{w} = [w_0, w_1, \ldots, w_N], \tag{1}$$

   onde N representa o número de entradas do perceptron.

2. Cálculo da saída estimada:

   Para cada exemplo de treinamento

   $$\mathbf{x}_j = [x_{j0}, \ldots, x_{jN}], \tag{2}$$

   calcula-se a saída estimada

   $$u_j = \mathbf{w} \cdot \mathbf{x}_j, \tag{3}$$

   isto é, um produto interno entre o vetor de pesos e o vetor de entradas.

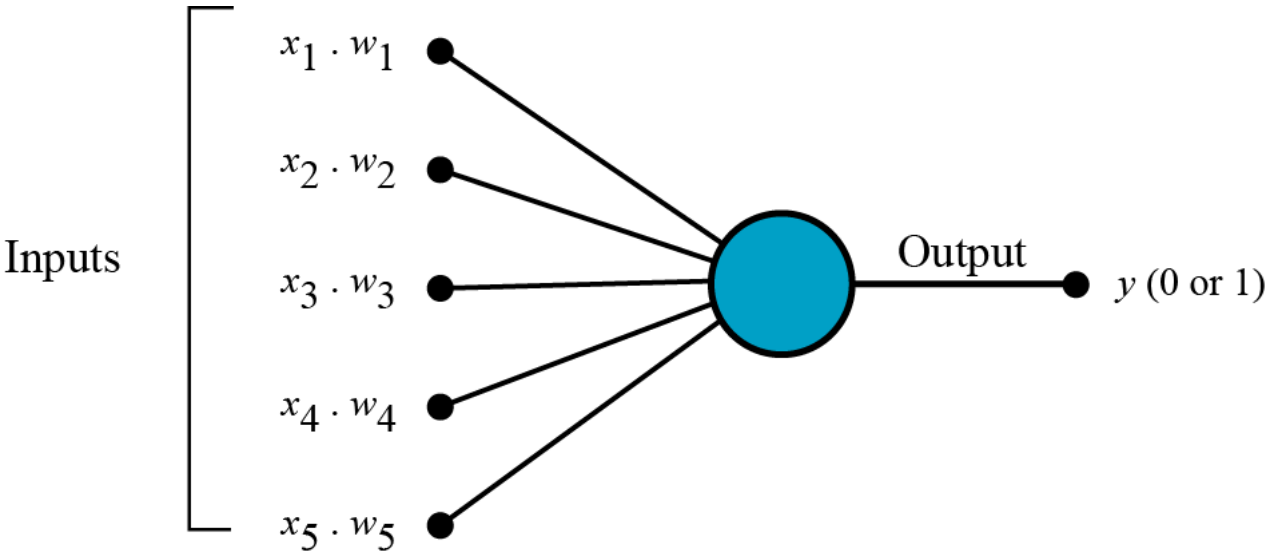

Figura 5 – Ilustração do Perceptron.

3. Atualização de saída:

   O valor de saída é ajustado com base em um função de ativação f, geralmente a sigmoide ou a função de sinal :

   $$y_j = f(u_j), \tag{4}$$

4. Cálculo do erro:

   O erro entre a sáida prevista yi e a saída desejada dj é calculado como:

   $$e_j = d_j - y_j, \tag{5}$$

5. Atualização dos pesos:

   O vetor de pesos é ajustado conforme a regra de aprendizado:

   $$\Delta \mathbf{w} = \eta \cdot e \cdot \mathbf{x}, \tag{6}$$

   onde é a taxa de aprendizado, parâmetro que controla o tamanho dos ajustes aplicados aos pesos.

6. Iteração:

   Os passos anteriores são repetidos até que seja atingido um número específico de épocas ou até que um critério de parada seja satisfeito.

A principal limitação do perceptron simples é a sua incapacidade de separar classes não linearmente separáveis, o que restringe significativamente seu campo de aplicação. Para superar essa limitação, surgiram as Redes Neurais Artificiais (RNAs), compostas por coleções organizadas de múltiplos neurônios artificiais interconectados.

Entre os modelos mais amplamente utilizados, destaca-se o Perceptron Multicamadas (MLP) que consiste em vários neurônios distribuídos em camadas interligadas. O sinal é propagado de forma unidimensional, da camada de entrada até a camada de saída, caracterizando uma arquitetura feedforward, ilustrada na Figura 6.

O treinamento do MLP é realizado por meio do algoritmo de retropropagação de erro (backpropagation), proposto por (BRYSON;HO,1969). Esse algoritmo busca minimizar o erro entre a saída produzida pela rede e a saída esperada, propagando o erro de forma retrógrada e ajustando os pesos de acordo com o gradiente do erro (POPESCU et al., 2009).

Os modelos MLP amplamente reconhecidos por sua capacidade de modelar relações não lineares complexas, apresentando bom desempenho em diferentes domínios e tamanhos de conjuntos de dados. Além disso, mantêm uma arquitetura relativamente simples quando comparados a redes neurais mais profundas.

Entretanto, algumas limitações devem ser consideradas:

- ❏ Baixa interpretabilidade, o que dificulta a compreensão do raciocínio subjacente às previsões realizadas pela rede;
- ❏ Elevado custo computacional, uma vez que o processo de treinamento de um MLP tende a ser demorado e requer recursos significativos, especialmente em conjuntos de dados de grande escala.

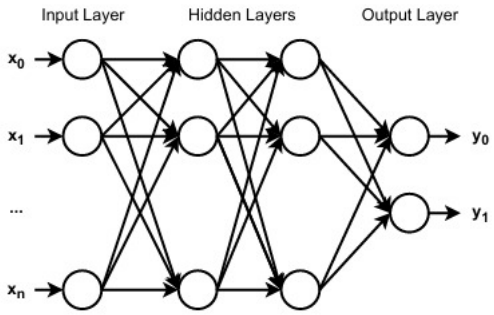

Figura 6 – Ilustração do MLP.

2.3 Previsão de Séries Temporais (TS)

Uma Série Temporal (TS) corresponde a um conjunto de dados obtidos a partir da observação contínua de um fenômeno ao longo do tempo. Os algoritmos de previsão de séries temporais têm como propósito modelar o comportamento temporal dessas observações, identificando padrões, tendências e sazionalidades nos dados históricos, a fim de realizar projeções futuras (PARMEZAN; SOUZA; BATISTA, 2019).

De forma geral, um algoritmo de previsão de séries temporais deve ser capaz de estimar valores futuros para um determinado horizonte de previsão (isto é, o número de períodos à frente que se deseja prever), com base em um conjunto de observações passadas, denominado janela de observação.

Os métodos utilizados para a previsão de séries temporais abrangem desde modelos estatísticos tradicionais, como o Autoregressivo Integrado de Médias Móveis (ARIMA) e o Autorregressivo Integrado de Médias Moveis Sazionais (SARIMA), até os modelos mais complexos baseados em Aprendizado de Máquina (AM), incluindo os algoritmos descritos nas seções subsequentes.

2.3.1 Rede Neural Recorrente (RNN)

As Redes Neurais Recorrentes (RNNs) diferem das redes neurais artificiais convencionais (ver Seção 2.2.3.1) por sua capacidade de reter informações ao longo do tempo. Em viturde dessa característica, costuma-se dizer que essas redes possuem "memória", pois utilizam informações provenientes de entradas anteriores para influenciar tanto as entradas atuais quanto as saídas subsequentes.

Essa propriedade é obtida por meio da presença de laços recorrentes nas conexões entre os neurônios, como ilustrado na Figura 7, e pela utilização de estados ocultos (hidden states). Esses estados são responsáveis por armazenar e transmitir informações históricas, permitindo que o modelo capture dependências temporais entre as sequências de dados desde os primeiros instantes até o momento atual.

Durante o processo de treinamento, as RNNs empregam uma variação do algoritmo de retroprogramação denomidada Retroprogramação no Tempo (Backpropramation Through Time - BPTT). No entanto, as RNNs tradicionais apresentam uma limitação significativa: sua dificuldade em reter informações de longo prazo.

Esse problema está associado ao fenômeno conhecido como desaparecimento do gradiente (vanishng gradients), no qual os gradientes calculados durante o treinamento diminuem progressivamente a cada iteração temporal, perdendo a capacidade de atualizar adequadamente os pesos das camadas iniciais da rede. Como consequência, o modelo torna-se incapaz de capturar relações de dependência de longo alcance.

Para contornar essa limitação, foram desenvolvidas arquiteturas recorrentes mais avançadas como as Redes de Memória de Longo e Curto Prazo (Long Short-Term Memory - LSTM) e as Unidades Recorrentes com Portas (Gated Recurrent Unit - GRU), ambas projetadas especificamente para preservar informações relevantes ao longo de períodos estendidos e mitigar o problema do desaparecimento do gradiente.

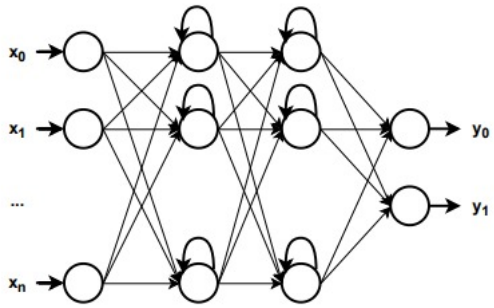

Figura 7 – Ilustração do RNN.

2.3.1.1 Memória de Curto e Longo Prazo (LSTM)

As Redes de Memória de Longo e Curto Prazo (Long Short-Term Memory - LSTM) introduzem o conceito de portões (gates), que são mecanismos internos responsáveis por regular o fluxo de informações dentro da rede. Esses portões controlam quando informações são adicionadas, mantidas ou removidas do estado da célula, permitindo que a rede preserve dependências temporais por períodos mais longos.

Cada portão da arquitetura LSTM utiliza funções de ativação específicas - tipicamente a sigmoide e a tangente hiperbólica (tanh) - para controlar o fluxo de dados.

A função sigmoide, que retorna valores entre 0 e 1, atua como um filtro que regula a proporção de informação transmitida ou bloqueada; já a função tanh é utilizada para sustentar o gradiente por mais tempo, constribuindo para mitigar o problema do desaparecimento do gradiente (vanishing gradients).

Uma arquiteruta LSTM simplificada é ilustrada na Figura 8 e é composta por três componentes principais:

- ❏ Portão de esquecimento (forget gate),
- ❏ Portão de entrada (input gate),
- ❏ Portão de saída (output gate).

O portão do esquecimento determina quanto da informação anterior deve ser retida ou descartada. Quando a saída da função sigmoide é 0, toda a informação é esquecida; quando é 1, toda a informação é mantida.

O portão de entrada quantifica a importância das novas informações apresentadas ao modelo, enquanto o portão de saída define o valor do próximo estado oculto (hidden state), que será transmitido à próxima célula da sequência temporal.

Desde sua proposição original em 1997, as LSTMs têm sido amplamente aplicadas em diferentes áreas de pesquisa, apresentando resultados expressivos em tarefas que envolvem dados temporais e dependências de longo prazo.

Por exemplo:

- ❏ (ABBES; MAGAGI; GOITA, 2019) empregaram uma LSTM para estimar a umidade do solo;
- ❏ (AKTER; LEE; KIM, 2021) aplicam LSTMs para previsão de consumo de energia, analisando 14 anos de dados horários;
- ❏ (P.; VANITHA; R, 2019) utilizou o modelo LSTM para previsão da velocidade do vento, com base em dados sazionais de quatro regiões da Índia.

Os resultados desses estudos demostram que as LSTMs são capazes de reduzir significativamente os erros de previsão em comparação com métodos convencionais de modelagem meteorológica, apresentando maior precisão nas estimativas e menor taxa de desvio nos valores previstos.

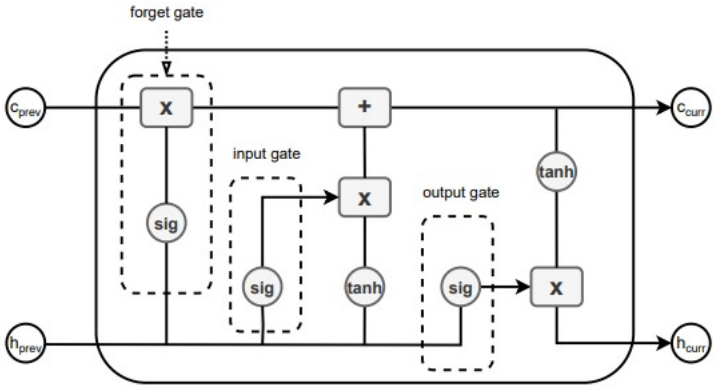

Figura 8 – Ilustração de um LSTM. c_{prev} representa o estado anterior da célula, h_{prev} representa o estado oculto anterior, c_{curr} representa o estado atual da célula e h_{curr} representa o estado oculto atual. Adaptado de (ZHAO et al., 2020).

2.3.1.2 Outras Arquiteturas Relacionadas

Outras arquiteturas também aparecem com frequência na literatura de previsão temporal, como GRU, CNN e modelos híbridos, por exemplo CNN-LSTM. A GRU busca resolver o problema do desaparecimento do gradiente com uma estrutura mais simples do que a LSTM, utilizando menos portões e, em muitos casos, menos custo computacional (ZHAO el al., 2020). Redes convulacionais também vêm sendo exploradas em séries temporais para extração de padrões locais e combinação com estruturas recorrentes (YAMASHITA et al,. 2018; MEHTAB; SEN; DASGUPTA, 2020; KOPRINSKA; WU; WANG, 2018). Entretanto, essas arquiteturas não foram implementadas neste trabalho e, por isso, são mencionadas apenas como referência de contexto, sem aprofundamento teórico detalhado.

2.3.2 Índices de Riscos Florestais

Os índices de risco de incêndios florestais constituem ferramentas importantes para o monitoramento e a prevenção de queimadas, pois permitem estimar, de forma quantitativa, condições ambientais favoráveis à ignição e à propagação do fogo. Em geral, esses índices são construídos a partir de variáveis meteorológicas, como precipitação, temperatura do ar, umidade relativa e, em alguns casos, velocidade do vento, sendo amplamente utilizados por órgãos ambientais e por estudos científicos como suporte ao planejamento de ações preventivas e operacionais.

Na literatura, diversos modelos foram propostos com esse propósito. Entre os índices abordados neste trabalho, destacam-se a Fórmula de Monte Alegre (FMA) e sua versão modificada, a FMA^{+}, além dos índices de Telicyn, Ångström e Nesterov. Cada um desses modelos foi desenvolvido com base em diferentes combinações de variáveis meteorológicas e apresenta critérios próprios para a classificação do risco de incêndio. Apesar dessas diferenças, todos compartilham o objetivo de sintetizar informações climáticas em indicadores capazes de auxiliar a identificação de períodos mais críticos para a ocorrência de fogo.

No contexto brasileiro, esses índices têm sido utilizados em diferentes estudos e aplicações operacionais, inclusive no bioma Pantanal. Nesse cenário, destaca-se o Sistema de Alerta de Risco de Incêndio do Pantanal (SARIPAN), que emprega variáveis meteorológicas e índices de risco como base para o monitoramento regional. Considerando essa relevância, as subseções seguintes apresentam a origem, a formulação e as principais características dos índices FMA, FMA^{+}, Telicyn, Ångström e Nesterov, bem como sua relação com sistemas de alerta aplicados ao contexto do Pantanal.

2.3.3 Fórmula de Monte Alegre (FMA) e Fórmula de Monte Alegre Modificada (FMA^{+})

A Fórmula de Monte Alegre (FMA) é o principal índice de risco de incêndios florestais desenvolvido com base nas condições climáticas do Brasil. O índice foi originalmente proposto por (SOARES, 1972), a partir de estudos realizados na região de Monte Alegre, localizada no centro do estado do Paraná.

O modelo leva em consideração variáveis meteorológicas como a umidade relativa do ar - medida às 13h - e a precipitação diária. Por se tratar de um índice acumulativo, seu valor depende da medição contínua dessas variáveis ao longo dos dias.

Quanto maior a sequência de dias com baixa umidade relativa e ausência de precipitação, maior será o risco de ocorrência de incêndios florestais.

A FMA enfatiza a probabilidade de ignição, isto é, a chance de início de um incêndio florestal.

Sua formulação é dada pela Equação 7:

$$FMA = \sum_{i=1}^{n} \frac{100}{H_i} \tag{7}$$

onde:

- n é o número de dias consecutivos sem precipitação (chuva inferior a 13 mm);
- H_i representa a umidade relativa do ar (%) medida às 13h.

O cálculo do índice está sujeito a ajustes diários baseados na precipitação, conforme indicado na Tabela 1.

Após o cálculo, o valor resultante é convertido em cinco classes de risco de incêndio florestal, apresentadas na Tabela 2, variando de baixo crítico.

Em (NUNES; SOARES; BATISTA, 2006), foi proposta a fórmula de Monte Alegre Modificada FMA^{+}, uma versão aprimorada da FMA original que incorpora a variável velocidade do vento ao cálculo do índice.

O objetivo da modificação foi desenvolver um modelo capaz de estimar não apenas a probabilidade de ignição, mas também o pontencial de propagação de incêndios florestais, ampliando a aplicabilidade do índice em diferentes condições meteorológicas.

Assim como a FMA, a FMA^{+} é um índice acumulativo, porém com sensibilidade adicional à dinâmica atmosférica, uma vez que a variável velocidade do vento, medida às 13h de cada dia, utilizada diretamente no cálculo.

Precipitação Diária (mm)	**Ajuste**
$\leq$ 2,4	Nenhum ajuste.
2,5 a 4,9	Subtrair 30% do valor da FMA calculado no dia anterior e adicionar $(100/H)$ do dia atual.
5,0 a 9,9	Subtrair 60% do valor da FMA calculado no dia anterior e adicionar $(100/H)$ do dia atual.
10,0 a 12,9	Subtrair 80% do valor da FMA calculado no dia anterior e adicionar $(100/H)$ do dia atual.
$>$ 12,9	Zerar o somatório (FMA = 0) e reiniciar o cálculo no dia seguinte.

Tabela 1 – Ajustes no cálculo do FMA conforme a precipitação diária. Adaptado de (SOARES; BATISTA, 2007).

FMA	**Risco de Incêndio Florestal**
$\leq$ 1,0	Nulo
1,1 a 3,0	Baixo
3,1 a 8,0	Médio
8,1 a 20,0	Alto
$>$ 20,0	Muito Alto

Tabela 2 – Classificação do risco de incêndio florestal com base na FMA, Adaptado de (SOARES; BATISTA, 2007)

.

O índice está sujeito aos mesmos ajustes de precipitação apresentados na Tabela 1 e pode ser calculado conforme a expressão da Equação 8:

$$FMA^{+} = \sum_{i=1}^{n} \left(\frac{100}{H_i} \right) \cdot e^{0,04 \cdot v_i}, \tag{8}$$

onde:

- n é o número de dias consecutivos sem chuva, isto é, com precipitação inferior a 13 milímetros;
- H_i representa a umidade relativa do ar(%), medida às 13h;
- v_i é a velocidade do vento (m/s), também medida às 13h.

A inclusão do termo exponencial $e^{0.04 \cdot v_i}$ permite ponderar o efeito da velocidade do vento sobre o risco de propagação do fogo, atribuindo maior peso a condições meteorológicas mais favoráveis à expansão das chamas.

Dessa forma, o FMA^{+} amplia a aplicabilidade da FMA original, tornando-se um indicador mais abrangente e sensível às variações climáticas locais.

O valor obtido por meio da Equação 8 deve ser classificado em uma das cinco categorias de risco de incêndio florestal, conforme definidos na tabela 3.

Essas classes representam diferentes níveis de propensão à ocorrência e propagação de incêndios, permitindo uma interpretação operacional dos resultados obtidos a partir do índice FMA^{+}.

FMA^{+}	**Risco de Incêndio Florestal**
$\leq$ 3,0	Nulo
3,1 a 8,0	Baixo
8,1 a 14,0	Médio
14,1 a 24,0	Alto
$>$ 24,0	Muito Alto

Tabela 3 – Classificação do risco de incêndio florestal com base na FMA^{+}. Adaptado de (SOARES;BATISTA, 2007).

2.3.4 Telicyn

O índice logarítmico de Telicyn foi proposto por (TELICYN, 1970), e seu cálculo é descrito pela Equação 9:

$$I = \sum_{i=1}^{n} \log(T_i - r_i), \tag{9}$$

onde:

- n é o número de dias consecutivos sem chuva, isto é, com precipitação inferior a 13 milímetros;
- T corresponde à temperatura do ar (°C), medida às 13h;
- r representa a temperatura do ponto de orvalho (°C), isto é, a temperatura em que o vapor da água presente na atmosfera se condensa, transformando-se em líquido.

De acordo com (SOARES; BATISTA, 2007), a temperatura do ponto de orvalho (r) pode ser obtida por meio de uma tabela de correlação entre a temperatura do ar e a umidade relativa.

O cálculo do índice está sujeito a ajustes baseados na ocorrência de precipitação, os quais são detalhados na Tabela 4.

Precipitação Diária (mm)	**Ajuste**
$> 2{,}5$	Interromper o somatório ($I = 0$) e reiniciar o cálculo no dia seguinte.

Tabela 4 – Ajustes no cálculo do índice Telicyn de acordo com a precipitação diária. Adaptado de (TORRES; RIBEIRO; 2008).

O valor obtido na Equação 9 deve ser convertido em uma das quatro categorias de risco de incêndio florestal,l conforme apresentado na Tabela 5.

Cabe destacar que, diferente dos índices FMA e FMA^{+} (ver Seção 2.3.4), o índice Telicyn não contempla a categoria de risco "Muito Alto", limitando-se, portanto, a quatro níveis distintos de classificação. Essa característica reflete a natureza mais restrita do modelo em relação à amplitude de variação das condições climáticas consideradas no cálculo.

Tabela 5 – Classes de Risco de Incêndio Florestal de acordo com o índice Telicyn.

I	**Risco de Incêndio Florestal**
$\leq 2{,}0$	Nulo
2,1 a 3,5	Baixo
3,6 a 5,0	Médio
$> 5{,}0$	Alto

Tabela 6 – Classes de Risco de Incêndio Florestal de acordo com o índice Telicyn. Adaptado de (ZICCARDI et al., 2020)

2.3.5 Angström

O índice de perigo de incêndio, também conhecido como fórmula de Ângström (B). foi desenvolvido na Suécia e é amplamente utilizado em diversas regiões da Escandinávia (ANGSTRÖM, 1942).

$$B = \left(\frac{H}{20}\right) + \left(\frac{T - 27}{10}\right), \tag{10}$$

onde:

- H representa a umidade relativa do ar (%), médida às 13h;
- T correnponde à temperatura do ar (°C), também medida às 13h.

O índice de Ângström apresenta uma estrutura simples e direta, pois não requer ajustes específicos e não é acumulativo, o que torna de fácil aplicação operacional em contextos de monitoramento climático.

O valor obtido a partir da Equação 10 deve, então, ser classificado em uma das quatro categorias de risco de incêndio florestal, conforme os intervalos apresentados na Tabela 6.

B	**Risco de Incêndio Florestal**
$< 3{,}5$	Nulo
3,5 a 3,9	Baixo
4,0 a 4,2	Médio
4,3 a 4,5	Alto
$> 4{,}5$	Muito Alto

Tabela 7 – Classes de risco de incêndio florestal de acordo com o índice Ângström. Adaptado de (CASAVECCHIA et al., 2019).

2.3.6 Nesterov

O índice de Nesterov foi originalmente desenvolvido na antiga União Soviética (NESTEROV, 1949) e baseia-se no conceito de inflamabilidade, que expressa a probabilidade de ocorrência de incêndios florestais a partir das condições meteorológicas.

Esse índice é calculado com base na soma dos "dias perigosos", isto é, dos dias caracterizados por elevado déficit de saturação do ar, conforme descrito na Equação 11:

$$G = \sum_{i=1}^{n} d_i \cdot T_i \tag{11}$$

onde:

- n representa o número de dias consecutivos sem chuva;
- d_i corresponde ao déficit de saturação (em milhares), medido às 13h;
- T_i é a temperatura do ar (°C), também medida às 13h.

O déficit de saturação (d) pode ser obtido por meio da Equação 12, que relaciona a pressão de saturação (E) e a umidade relativa do ar (H):

$$d = E \cdot \left(1 - \frac{H}{100}\right), \tag{12}$$

onde:

- E representa a pressão máxima de vapor (em milbares);
- H corresponde à umidade relativa do ar (%).

O índice de Nesterov, portanto, integra informações de temperatura e umidade relativa para expressar o acúmulo de calor e secura atmosférica ao longo dos dias, tornando-se um dos indicadores mais utilizados internacionalmente para o monitoramento do risco de incêndios florestais.

De acordo com (SOARES; BATISTA, 2007), o valor de E pode ser obtido por meio de uma tabela de correlação entre a temperatura do ar e a pressão máxima de vapor.

O cálculo do índice está sujeito a ajustes em função da precipitação, uma vez que esta reduz a inflamabilidade da vegetação e, consequentemente, o risco de incêndios florestais.

Esses ajustes estão escritos na Tabela 7.

Precipitação Diária (mm)	**Ajuste**
$\leq$ 2,0	Nenhum.
2,1 a 5,0	Subtrair 25% do valor de G calculado no dia anterior e adicionar $(d \cdot t)$ do dia atual.
5,1 a 8,0	Subtrair 50% do valor de G calculado no dia anterior e adicionar $(d \cdot t)$ do dia atual.
8,1 a 10,0	Descartar a soma anterior de G e iniciar um novo cálculo. Ou seja, $G = (d \cdot t)$ do dia atual.
$>$ 10,1	Interromper o somatório ($G = 0$) e reiniciar o cálculo no dia seguinte.

Tabela 8 – Ajustes no cálculo do índice de Nesterov conforme a precipitação. Adaptado de (SOARES; BATISTA, 2007).

O valor obtido por meio da Equação 25 deve então ser convertido em uma das cinco classes de risco de incêndio florestal, conforme mostrado na Tabela 8.

G	**Risco de Incêndio Florestal**
$\leq$ 300	Nulo
301 a 500	Baixo
501 a 1000	Médio
1001 a 4000	Alto
$>$ 4000	Muito Alto

Tabela 9 – Classes de risco de incêndio florestal de acordo com o índice de Nesterov.Adaptado de (SOARES; BATISTA, 2007)

2.3.7 Comparação dos Índices

Os índices de risco de incêndio florestal apresentados nas seções anteriores deste capítulo baseiam-se em uma ou mais das seguintes variáveis climáticas primárias: precipitação (mm), umidade relativa do ar (%) temperatura do ar (°C) e velocidade do vento (m/s).

A relação entre essas variáveis e os respectivos índices é sintetizada na Tabela 9.

Tabela 10 – Variáveis climáticas exigidas para cada um dos principais índices de risco de incêndio florestal.

Índice	**Precipitação**	**Umidade Relativa**	**Temperatura**	**Velocidade do Vento**
FMA	X			
FMA^+	X	X		X
Telicyn	X		X	
Ângström		X	X	
Nesterov	X	X	X	

Tabela 11 – Variáveis climáticas exigidas para cada um dos principais índices de risco de incêndio florestal.

Além desses índices, a literatura especializada apresenta outras formulações complementares, que podem incorporar variáveis climáticas adicionais para representar de forma mais ampla as condições de risco. Entre elas, destacam-se o índice P-EVAP e EVAP/P (SAMPAIO, 1991) ambos fundamentados na relação entre precipitação (P) e evaporação (EVEAP), expressas em milímetros (mm).

Outro índice amplamente empregado em diferentes regiões do mundo é o Fire Weather Index (FWI), desenvolvido no Canadá na década de 1970 e continuamente aperfeiçoado desde então (WAGNER, 1987). O FWI é composto por diversos subíndices que representam diferentes aspectos das condições ambientais, incluindo o teor de umidade da camada orgânica do solo eo índice de seca, que reflete o défict de umidade.

Essa estrutura hierárquica confere ao FWI uma maior sensibilidade às variações meteorológicas, tornando-o uma das ferramentas mais completas para a avaliação do risco de incêndios florestais.

Em (TORRES; RIBEIRO, 2008), os índices FMA, FMA^+, P-EAV/P foram aplicados a dados meteorológicos da região de Juiz de Fora (MG, Brasil).

Os dados foram coletados em diferentes horários do dia, e observou-se que as melhores correlações ocorreram quando a umidade relativa e a temperatura do ar foram medidas às 13h.

Os resultados indicaram que tais índices apresentam maior eficiência na previsão de ocorrência de incêndios florestais, em comparação a simples detecção de eventos.

Entre os índices avaliados, o EVAP/P destacou-se por apresentar melhor desempenho geral.

De forma semelhante, (TORRES et al., 2017) aplicaram os índices FMA, FMA^+, P-EAV/P, EVAP/P, FWI, Nesterov e Telicyn a dados da região de Viçosa (MG, Brasil).

Os resultados mostraram que o índice de Telicyn foi o mais eficiente para o contexto analisado, pelos índices EVAP/P e P-EVAP, evidenciando a importância de variáveis de défict hídrico na previsão do risco de incêndio.

No estudo de (TORRES; LIMA, 2019) os índices FMA, FMA$^+$, Nesterov, Telicyn, P-EVEAP, EVA/P e FWI foram avaliados com dados da Serra do Brigadeiro (MG, Brasil).

Os autores identificaram que os índices P-EVAP e FWI apresentam melhor desempenho na previsão da ocorrência de incêndios florestais, reforçando a utilidade desses modelos em regiões com forte variabilidade climática.

Especialmente na região do Pantanal brasileiro, (SORIANO; DANIEL; SANTOS, 2015) conduziram um estudo comparativo utilizando dados do Pantanal sul-mato-grossense.

Foram avaliados os índices FMA, FMA$^+$, Nesterov, Telicyn e Ångström.

Os resultados indicaram que, para a detecção das classes de alto risco ("Muito Alto"e "Alto"), no período de 1999 a 2008, o índice Nesterov apresentou maior eficiência, seguido pelo FMA.

Contudo, ao considerar todas as classes de risco, o FMA obterve a maior acurácia geral.

Ainda no contexto pantaneiro, destaca-se o Sitema de Alerta de Risco de Incêndio do Pantanal (SARIPAN), desenvolvido por (NARCISO; SORIANO, 2019).

O SARIPAN utiliza os índices FMA, FMA$^+$, Nesterov, Telicyn e Ångström para identificar áreas com risco de fogo em tempo quase real.

O sistema é capaz de gerar previsões apenas para o mesmo dia, a partir das variáveis meteorológicas observadas, e não realiza previsões multidiárias.

Além disso, o SARIPAN não incorpora técnicas de Aprendizado de Máquina (AM), o que limita sua capacidade de aprender com dados históricos e adaptar-se às mudanças climáticas.

Por essas razões, os índices FMA, FMA$^+$, Nesterov, Telicyn e Ångström, bem como o SARIPAN, são todos baseados em medições pontuais de variáveis climáticas - precipitação, umidade relativa, temperatura do ar e velocidade do vento -, conforme descrito no capítulo posterior a este.

Desta forma, apenas o SAPIRAN foi incluído na comparação com os resultados obtidos pelo sistema desenvolvido neste estudo.

2.3.8 Medidas Preventivas

De acordo com (SOARES; BATISTA, 2007), diferentes medidas preventivas devem ser adotadas conforme o nível de risco de incêndio florestal, indicado pelos índices apresentados nas seções anteriores.

Essas medidas visam tanto à redução da probabilidade de ignição quando à mitigação dos impactos em caso de ocorrência de incêndios.

As recomendações para cada classe de risco são descritas a seguir :

- Risco Nulo:

  Nesse nível, considera-se que não há risco de ocorrência de incêndios florestais.

  O período é ideal para atividades de capacitação e planejamento, incluindo o treinamento de equipes, a manuntenção preventiva e a revisão de equipamentos.

  A vigilância preventiva pode ser temporariamente desmobilizada, e as torres de comando e observação não precisam permanecer operacionais.

- Risco Baixo:

  Há possibilidade de ocorrência de incêndios, embora o risco seja considerado reduzido.

  Esse período deve ser aproveitado para intensificar o treinamento de pessoal, planejar ações preventivas e realizar manutenções de rotina.

  A vigilância pode ser mantida em nível mínimo, e as torres de comando ainda não necessitam operar em tempo integral.

- Risco Médio:

  O risco de incêndios florestais é moderado, exigindo atenção operacional.

  Devem ser preparadas as equipes de combate e os equipamentos de suporte, que devem permanecer em estado de alerta.

  Veículos e equipamentos de comunicação devem ser acionados e testados diariamente, e as torres de vigilância devem iniciar suas atividades operacionais.

- Risco Alto:

  Nesse nível, o risco de incêndios é significativo, demandando estado de prontidão por parte das equipes de combate e dos equipamentos.

  As atividades agrícolas e florestais que envolvem o uso de fogo devem ser monitoradas e, quando necessário, restringidas.

  Os equipamentos de comunicação e veículos devem ser testados ao menos duas vezes por dia, e a vigilância preventiva deve ser intensificada, com ampliação do horário de funcionamento das torres de comando e observação.

- Risco Muito Alto:

  O risco é extremamente elevado, e a mobilização total das equipes de combate torna-se obrigatória.

  As atividades agrícolas e florestais que envolvem fogo devem ser imediatamente suspensas.

  A população local deve ser alertada por meio dos canais oficiais de comunicação, e as equipes de primeira resposta devem permanecer em prontidão máxima.

A vigilância preventiva deve ser reforçada, com operação contínua das torres de observação e comando.

Essas diretrizes constituem um protocolo de ação progressiva, no qual o nível de resposta preventiva aumenta proporcionalmente ao risco identificado pelos índices de perigo de incêndio florestal.

3 Metodologia e Desenvolvimento do Projeto

Conforme apresentado no Capítulo 1, este estudo tem como objetivo analisar e prever o risco de incêndios florestais no Pantanal brasileiro, a partir do histórico de focos de incêndio ativos detectados por satélites de referência, agregados em base mensal, no período compreendido entre 1998 e 26 de janeiro de 2026. Com base nessa série histórica, busca-se identificar padrões temporais e tendências, de modo a viabilizar previsões para meses e anos futuros, contribuindo para o planejamento estratégico e a adoção de medidas preventivas por parte dos órgãos competentes.

Para atingir esse objetivo, serão utilizados dois modelos distintos para fins comparativos: o XGBoost e a Rede Neural do tipo LSTM (Long Short-Term Memory), ambos aplicados aos dados de focos de incêndio ativos registrados no bioma Pantanal.

3.1 Justificativa Formal da Escolha dos Modelos

Embora a literatura apresente diversas abordagens para previsão de séries temporais — incluindo modelos estatísticos clássicos, como ARIMA e SARIMA, bem como métodos supervisionados tradicionais, como SVM e KNN — este estudo optou por concentrar a análise comparativa em dois modelos representativos de paradigmas distintos de modelagem preditiva: o XGBoost e a rede neural recorrente do tipo Long Short-Term Memory (LSTM).

A escolha desses modelos foi metodologicamente intencional, fundamentada na necessidade de comparar duas estratégias conceitualmente distintas de tratamento da dependência temporal:

- ❏ um modelo baseado em árvores de decisão e técnica de ensemble (gradient boosting);
- ❏ um modelo baseado em redes neurais recorrentes profundas, projetadas especificamente para dados sequenciais.

Essa comparação permite avaliar o impacto do tratamento explícito versus implícito da estrutura temporal no desempenho preditivo aplicado à série histórica de focos de incêndio no Pantanal brasileiro.

3.1.1 Justificativa da Escolha do XGBoost

O XGBoost foi selecionado por apresentar características particularmente relevantes para bases tabulares derivadas de séries temporais. Trata-se de um algoritmo de gradient

boosting baseado em árvores de decisão, amplamente reconhecido por seu desempenho em problemas de regressão e classificação.

Entre as principais razões para sua escolha, destacam-se:

- ❏ Capacidade de modelar relações não lineares complexas;
- ❏ Robustez a ruído e valores extremos;
- ❏ Presença de mecanismos de regularização (penalizações L1 e L2), que reduzem o risco de overfitting;
- ❏ Desempenho consistente mesmo com volume moderado de dados;
- ❏ Eficiência computacional, permitindo execução em ambientes com infraestrutura limitada.

Embora o XGBoost não modele explicitamente dependências temporais, sua aplicação torna-se viável por meio da engenharia de atributos temporais, como defasagens (lags), médias móveis, estatísticas de volatilidade e variáveis sazonais. Dessa forma, o problema de previsão temporal é reformulado como um problema de regressão supervisionada com atributos derivados.

3.1.2 Justificativa da Escolha da LSTM

A rede neural LSTM foi escolhida por ser uma arquitetura desenvolvida especificamente para modelagem sequências temporais. Diferentemente de modelos baseados em árvores, a LSTM incorpora mecanismos internos de memória, denominado gates, que regulam o fluxo de informação ao longo da sequência.

Seu diferencial reside na capacidade de:

- ❏ Capturar dependências de curto e longo prazo;
- ❏ Modelar padrões sazonais recorrentes;
- ❏ Aprender representações temporais diretamente da sequência de dados.

Ao contrário do XGBoost, a LSTM não depende da criação manual de atributos temporais, pois sua própria estrutura é projetada para extrair automaticamente dependências ao longo do tempo. Essa característica torna a LSTM particularmente adequada para avaliar o potencial de aprendizado implícito da estrutura temporal presente na série histórica.

3.1.3 Fundamentação Comparativa

A comparação entre XGBoost e LSTM permite analisar duas filosofias distintas de aprendizado aplicadas ao mesmo problema:

- ❏ Modelagem baseada em partições hierárquicas do espaço de atributos, característica de algoritmos baseados em árvores;
- ❏ Modelagem baseada em estados internos e memória temporal, característica de redes neurais recorrentes.

Essa análise é especialmente relevante no contexto ambiental, no qual séries temporais frequentemente apresentam sazonalidades bem definida, tendências interanuais e ocorrência de eventos extremos episódicos.

Ao comparar esses dois paradigmas, busca-se não apenas identificar o modelo com melhor desempenho preditivo, mas também compreender como diferentes estratégias de tratamento da dependência temporal influenciam a capacidade de generalização em cenários caracterizados por alta variabilidade e assimetria.

3.2 Variáveis de Observações

Os dados utilizados neste estudo foram obtidos na plataforma TerraBrasilis, do Programa Queimadas do Instituto Nacional de Pesquisas Espaciais (INPE), a partir da base de monitoramento de focos ativos por estado, região ou bioma, disponível para consulta e download em formato CSV (INPE, ano).

O conjunto de dados de variáveis observacionais consiste em uma série temporal compreendendo o período de 1998 até o início de 2026, contendo registros de focos ativos de incêndio florestal na região do Pantanal brasileiro.

Esses dados são provenientes do Programa Queimadas, mantido pelo Instituto Nacional de Pesquisas Espaciais (INPE), e são disponibilizados publicamente, garantindo transparência e reprodutibilidade científica.

A Figura 9 apresenta um resumo descritivo do conjunto de dados, evidenciando a distribuição temporal dos focos ativos de incêndio ao longo do período analisado.

Ano	Janeiro	Fevereiro	Março	Abril	Maio	Junho	Julho	Agosto	Setembro	Outubro	Novembro	Dezembro	Total
1998	-	-	-	-	-	12	206	172	542	507	162	58	1659
1999	161	10	17	29	106	65	845	2988	1922	2049	743	52	8987
2000	93	30	19	1	8	29	25	263	503	950	295	74	2290
2001	66	62	11	36	67	219	444	2540	2264	879	175	19	6782
2002	137	28	10	24	29	320	805	2934	2662	2761	2328	448	12486
2003	180	48	36	40	107	170	260	517	1188	715	378	83	3722
2004	185	126	141	68	15	240	384	1164	3963	1912	335	155	8688
2005	20	107	100	163	331	435	1259	5993	2997	933	125	73	12536
2006	28	51	36	19	104	101	375	892	1024	266	254	23	3173
2007	6	13	30	68	101	239	341	1858	5498	1481	189	45	9869
2008	14	13	20	28	48	44	216	588	1660	1046	274	594	4545
2009	380	117	71	525	815	308	311	695	1127	919	414	55	5737
2010	31	47	88	87	67	247	511	1548	3072	1142	385	795	8020
2011	145	22	2	2	20	42	105	309	807	562	873	643	3532
2012	188	83	97	38	115	109	490	2698	2518	832	157	122	7447
2013	108	55	115	51	47	17	129	440	1201	544	513	176	3396
2014	103	64	23	55	16	27	90	134	375	459	184	37	1567
2015	95	51	28	29	36	218	225	1025	1181	794	282	494	4458
2016	37	29	18	34	59	93	542	966	2000	1066	215	125	5184
2017	261	73	68	38	48	93	610	1092	2588	669	214	19	5773
2018	23	8	14	19	28	46	190	275	785	120	20	163	1691
2019	337	211	93	33	68	239	494	1690	2887	2430	1296	247	10025
2020	265	164	602	784	313	406	1684	5935	8106	2856	778	223	22116
2021	41	35	50	87	60	98	508	1505	2954	2515	257	148	8258
2022	83	61	45	25	188	115	294	96	242	48	201	239	1637
2023	21	9	12	15	33	77	126	110	373	1157	4134	513	6580
2024	310	73	176	94	246	2639	1218	4411	2688	2437	191	15	14498
2025	34	15	4	7	10	16	39	48	189	241	46	75	724
2026	74	-	-	-	-	-	-	-	-	-	-	-	74
Máximo*	380	211	602	784	815	2639	1684	5993	8106	2856	4134	795	22116
Média*	124	59	71	89	114	238	454	1532	2047	1153	551	204	6621
Mínimo*	6	8	2	1	8	12	25	48	189	48	20	15	724

Figura 9 – Comparação do total de focos ativos detectados pelo satélite de referência em cada mês, no período de 1998 até 26/jan.

3.3 Estratégia de Divisão Temporal e Validação de Dados

Por se tratar de um problema de previsão de séries temporais, a divisão dos dados foi realizada respeitando rigorosamente a ordem cronológica dos eventos, evitando qualquer forma de vazamento de informação (data leakage).

Diferentemente de problemas tradicionais de aprendizado supervisionado, nos quais é comum utilizar validação cruzada aleatória, em séries temporais essa prática comprometeria a integridade do experimento, pois permitiria que informações futuras influenciassem o treinamento.

3.3.1 Divisão para o XGBoost

No caso do modelo XGBoost, a base foi dividida em:

- ❏ Conjunto de treinamento: período inicial da série;
- ❏ Conjunto de validação: subconjunto do período de treinamento utilizado para aplicação de early stopping;
- ❏ Conjunto de testes: período final da série, completamente isolado durante o treinamento.

A separação foi realizada preservando a sequência temporal dos dados, garantindo que o modelo fosse treinado exclusivamente com informações passadas para prever meses futuros.

O early stopping foi aplicado em base no desempenho no conjunto de validação, interrompendo o treinamento quando não houve melhoria significativa na métrica do erro, prevenindo sobreajuste.

3.3.2 Divisão para LSTM

Para LSTM, o procedimento seguiu a mesma lógica temporal. Após a criação das sequências com janela deslizante = (12 meses), as amostras foram separadas cronologicamente em:

- ❏ Conjunto de treino;
- ❏ Conjunto de validação (utilizado para monitoramento da função perda);
- ❏ Conjunto de teste.

O conjunto de teste foi mantido completamente isolado durante o treinamento da rede.

O uso de early stopping e redução adaptativa da taxa de aprendizado (ReduceLROnPlateau) foi baseado exclusivamente no desempenho no conjunto de validação, assegurando que o modelo não tivesse acesso indireto aos dados de teste.

3.3.3 Garantia de Integridade Temporal

Em ambos os modelos:

- ❏ Não houve embaralhamento aleatório (shuffle) dos dados;
- ❏ A sequência cronológica foi mantida;
- ❏ As previsões futuras foram realizadas por abordagem recursiva, utilizando apenas informações disponíveis no instante anterior.

Os procedimentos adotados garantem que a comparação entre XGBoost e LSTM seja metodologicamente justa, pois ambos foram avaliados sob a mesma estrutura temporal, como divisão cronológica consistente e ausência de vazamento de informação

3.4 Implementação XGBoost

A implementação do modelo XGBoost foi realizada em um ambiente computacional controlado, com o objetivo de assegurar a reprodutibilidade dos experimentos, a rastreabilidade das etapas de processamento e a coerência metodológica com a abordagem de previsão temporal adotada neste estudo. O modelo foi estruturado segundo uma arquitetura modular, permitindo a separação lógica entre as etapas de carregamento dos dados, engenharia de atributos, pré-processamento, treinamento e avaliação do desempenho preditivo. Para o treinamento do modelo, utilizou-se a biblioteca XGBoost, amplamente empregada em aplicações baseadas em gradient boosting sobre árvores de decisão.

O processamento foi executado em um ambiente computacional padrão, utilizando exclusivamente recursos de CPU, sem o uso de aceleração por GPU. Essa decisão deve-se ao fato de que o volume de dados e a complexidade do modelo mostraram-se plenamente compatíveis com o processamento em CPU, sem prejuízo ao desempenho ou à estabilidade do treinamento. Além disso, essa escolha reforça a viabilidade prática da aplicação do método em contextos institucionais com infraestrutura computacional limitada, ampliando o potencial de replicação e adoção da abordagem proposta por órgãos públicos e instituições de pesquisa.

3.4.1 Organização do Pipeline de Processamento

O pipeline de modelagem foi estruturado de forma sequencial e objetiva, contemplando as seguintes principais etapas:

1. Carregamento e validação dos dados provenientes do Programa Queimadas (INPE).
2. Transformação da base para formato temporal longo, com agregação mensal.
3. Engenharia de atributos temporais, incluindo defasagens (lags), média móveis, estatísticas rolantes e variáveis sazonais;
4. Normalização de variáveis explicativas, assegurando a estabilidade numérica durante o treinamento;
5. Divisão temporal dos dados em conjunto de treino e teste, respeitando a ordem cronológica dos fatos;
6. Treinamento do modelo XGBoost, com estratégia de early stopping;

7. Avaliação do desempenho preditivo por meio de métricas estatísticas;

8. Geração de previsões futuras, utilizando abordagem recursiva (multi-step forecast).

Essa organização garante que o modelo não incorra em vazamento de informação temporal (data leakage), aspecto crítico em problemas de previsão de séries temporais.

3.4.2 Formulação Conceitual do Modelo XGBoost no Contexto Temporal

Embora o XGBoost não seja um modelo temporal explícito, sua aplicação a séries temporais é viabilizada por meio da incorporação explícita da estrutura temporal nos atributos de entrada. Assim, o problema de previsão é reformulado como uma tarefa de regressão supervisionada, em que o valor de focos de incêndio em um determinado mês é estimado a partir de observações passadas.

Formalmente, o modelo busca aprender a função:

$$\hat{y}_t = f(X_t)$$

em que:

- $\hat{y}_t$ representa o número previsto de focos de incêndio no mês t;
- X_t é o vetor de atributos construído a partir do histórico da série temporal até o instante $t-1$;
- $f(\cdot)$ corresponde ao ensemble de árvores de decisão treinado via gradient boosting.

No contexto do XGBoost, essa função é expressa com a soma de K árvores de decisão:

$$\hat{y}_t = \sum_{k=1}^{K} f_k(X_t)$$

onde cada f_k é ajustada de forma iterativa para corrigir os erros residuais das árvores anteriores.

3.4.3 Estratégia de Previsão Futura (Multi-step Forecast)

Para a projeção de valores futuros, foi adotada uma abordagem de previsão recursiva (recursive forecasting). Nessa estratégia, a previsão obtida para um determinado mês passa a ser utilizada como entrada na construção das variáveis explicativas do mês subsequente, permitindo a extensão do horizonte de previsão para períodos mais longos, como cinco anos à frente.

Essa metodologia mostra-se particularmente adequada em cenários nos quais não há disponibilidade de informações exógenas futuras, como projeções climáticas ou variáveis

ambientais previstas. Dessa forma, a abordagem é consistente com o uso exclusivo do histórico observado de focos de incêndio, garantindo coerência metodológica com os dados efetivamente disponíveis ao longo da série temporal.

A seguir, apresenta-se o algoritmo que sintetiza o procedimento adotado nesta seção, descrevendo as etapas de implementação do modelo XGBoost, desde o uso dos dados históricos até o processo de previsão recursiva para valores futuros. Esse algoritmo tem como objetivo formalizar metodologicamente as decisões tomadas durante o desenvolvimento do modelo e facilitar a reprodutibilidade do estudo.

Algoritmo 1 Pipeline de Previsão de Focos de Incêndio com XGBoost

Require: Série temporal mensal de focos de incêndio $Y = \{y_1, y_2, \ldots, y_T\}$
Require: Horizonte de previsão H
Ensure: Previsões futuras $\hat{Y} = \{\hat{y}_{T+1}, \hat{y}_{T+2}, \ldots, \hat{y}_{T+H}\}$
Carregar dados históricos
Validar e transformar dados para formato temporal
Construir variáveis temporais (mês, estação, seno/cosseno)
Construir defasagens (*lags*) e estatísticas móveis
Normalizar variáveis explicativas
Dividir dados em treino e teste respeitando a ordem temporal
Treinar o modelo XGBoost com *early stopping*
Avaliar desempenho com métricas estatísticas
for $h = 1$ **to** H **do**
    Construir vetor de atributos X_{T+h} a partir do histórico disponível
    $\hat{y}_{T+h} \leftarrow f(X_{T+h})$
    Atualizar histórico com $\hat{y}_{T+h}$
end for
return $\hat{Y}$

3.5 Implementação LSTM

A implementação do modelo LSTM (Long Short-Term Memory) foi realizada em um ambiente computacional controlado, com o objetivo de assegurar a reprodutibilidade dos experimentos, a rastreabilidade do pipeline de processamento e a coerência metodológica com a tarefa de previsão temporal proposta neste estudo.

O fluxo de implementação foi estruturado de forma modular, permitindo a separação lógica entre as etapas de carregamento e validação dos dados, transformação da base em série temporal, pré-processamento, geração de sequências temporais, treinamento do modelo, avaliação do desempenho e projeção de valores futuros.

Essa organização modular facilita tanto a manutenção do código quanto a replicação dos experimentos, além de possibilitar ajustes independentes em cada etapa do processo, sem comprometer a integridade do pipeline como um todo.

O processamento foi realizado em um ambiente computacional padrão, utilizando exclusivamente recursos de CPU, sem o uso de aceleração por GPU. Embora o modelo

computacional implementado neste estudo seja compatível com o uso desse recurso adicional, o volume de dados disponível — correspondente a uma série temporal mensal — não demanda tal capacidade computacional para um processamento eficiente.

Além disso, tanto o tamanho do conjunto de dados quanto o porte do modelo adotado, composto por duas camadas LSTM com 50 unidades cada, mostraram-se plenamente compatíveis com a execução em CPU. Essa característica possibilita que o experimento seja facilmente replicado em ambientes com infraestrutura computacional limitada, ampliando o potencial de aplicação prática da abordagem proposta.

3.5.1 Organização do Pipeline de Processamento

O pipeline de modelagem com LSTM foi estruturado de uma forma sequencial e objetiva, contemplando as seguintes etapas principais:

1. Carregamento dos dados a partir de arquivo Excel (base histórica mensal).
2. Validação estrutural (coluna "Ano" e meses esperados) e filtragem de linhas inválidas (por exemplo, totais ou resumos do Excel).
3. Conversão para formato longo (Ano–Mês–Focos) e criação da coluna temporal (*Data*).
4. Tratamento de valores faltantes no final da série, cortando o conjunto de dados no último mês com valor observado, evitando a inserção de "meses futuros" como dados reais.
5. Visualização exploratória, incluindo análise da série temporal e análise sazonal (média mensal, boxplot mensal, total anual e heatmap ano–mês).
6. Normalização da variável alvo por meio do *MinMaxScaler*, garantindo estabilidade numérica durante o treinamento da rede.
7. Criação de sequências temporais com janela deslizante ($lookback = 12$), definindo o problema como uma previsão de um passo à frente.
8. Divisão temporal dos dados em conjuntos de treino e teste, preservando a ordem cronológica e evitando *data leakage*.
9. Treinamento do modelo LSTM com regularização (*dropout*), ajuste da taxa de aprendizado (*ReduceLROnPlateau*) e aplicação de *early stopping*.
10. Avaliação do desempenho preditivo por meio de métricas estatísticas (RMSE, MAE, R^2, MAPE) e análise de resíduos (posteriormente apresentados no capítulo seguinte).

11. Geração de previsões futuras por estratégia recursiva (*multi-step forecast*), em que cada previsão alimenta a etapa subsequente.

Essa organização mantém a integridade temporal do problema, evitando vazamentos de informação e garantindo que o modelo seja avaliado em uma simulação realista de previsão futura.

3.5.2 Formulação Conceitual do Modelo LSTM no Contexto Temporal

Diferentemente de métodos baseados em árvores de decisão, como o XGBoost, o modelo LSTM é uma rede neural recorrente projetada especificamente para aprender dependências temporais diretamente a partir da estrutura sequencial dos dados. Nesse contexto, a dinâmica temporal não precisa ser explicitamente codificada por meio de atributos derivados, uma vez que o próprio modelo é capaz de capturar relações de curto e longo prazo ao longo da série histórica.

Para possibilitar o uso da LSTM, a série temporal é reformulada como um problema de aprendizado supervisionado por meio da construção de sequências temporais. Nessa abordagem, o vetor de entrada é composto por uma janela deslizante contendo os valores observados nos últimos L meses, enquanto o valor-alvo corresponde ao mês imediatamente subsequente.

Formalmente, considera-se uma série temporal mensal $Y = \{y_1, y_2, \ldots, y_T\}$. Para cada instante t, define-se a sequência de entrada:

$$x_t = [y_{t-L},\, y_{t-L+1},\, \ldots,\, y_{t-1}]$$

e o alvo :

$$\hat{y}_t = f_\theta(\mathbf{x}_t)$$

em que $f\theta(.)$ representa a função aprendida pela rede LSTM parametrizada por θ. Na prática, o seu pipeline utiliza $L = 12$, de modo que o modelo tenta capturar padrões de sazonalidade anual (ciclo de 12 meses) e tendências de curto prazo.

3.5.3 Estratégia de Previsão Futura (Multi-step Forecast)

Para a projeção de valores futuros, foi adotada uma estratégia de previsão recursiva (recursive forecasting). Nessa abordagem, a previsão obtida para o próximo mês é incorporada à sequência de entrada, substituindo a observação real — inexistente em cenários futuros —, e o modelo é então executado novamente para estimar o valor do mês subsequente. Esse processo é repetido iterativamente até que o horizonte de previsão desejado seja alcançado.

Esse procedimento mostra-se especialmente adequado em situações nas quais não há disponibilidade de variáveis exógenas futuras, como projeções de clima, vento ou precipitação. Dessa forma, a metodologia permanece coerente com a disponibilidade real dos dados e com o uso exclusivo do histórico observado de focos de incêndio, assegurando consistência metodológica ao processo de previsão.

A seguir, apresenta-se o algoritmo que sintetiza o procedimento adotado nesta seção, descrevendo de forma estruturada as etapas de implementação do modelo LSTM. O algoritmo abrange desde o uso e preparação dos dados históricos, passando pela construção das sequências temporais e pelo treinamento do modelo, até o processo de previsão recursiva para valores futuros. Essa formalização tem como objetivo sistematizar a metodologia empregada e facilitar a reprodutibilidade dos experimentos.

Algoritmo 2 Pipeline de Previsão Mensal de Focos de Incêndio com LSTM

Require: Série temporal mensal de focos de incêndio $Y = \{y_1, y_2, \ldots, y_T\}$
Require: Janela temporal (*lookback*) L e horizonte H
Ensure: Previsões futuras $\hat{Y} = \{\hat{y}_{T+1}, \hat{y}_{T+2}, \ldots, \hat{y}_{T+H}\}$
Carregar base histórica (Excel) e validar estrutura (Ano e meses)
Remover linhas inválidas e transformar para formato longo
Criar coluna temporal (Data) e ordenar cronologicamente
Cortar meses finais sem observação real (tratamento de *missing values*)
Gerar visualizações exploratórias
Normalizar Y via MinMaxScaler: $Y' \leftarrow scale(Y)$
Construção do dataset supervisionado:
for $t = L + 1$ **to** T **do**
    $\mathbf{x}_t \leftarrow [y'_{t-L}, \ldots, y'_{t-1}]$
    $z_t \leftarrow y'_t$
end for
Dividir em treino e teste (respeitando ordem temporal)
Treinar LSTM com regularização (*dropout*) e *early stopping*
Gerar previsões no teste: $\hat{z}_t \leftarrow f_\theta(\mathbf{x}_t)$
Desnormalizar: $\hat{y}_t \leftarrow inverse\_scale(\hat{z}_t)$
Avaliar desempenho (RMSE, MAE, MAPE) e analisar resíduos
Previsão futura recursiva:
$\mathbf{s} \leftarrow [y'_{T-L+1}, \ldots, y'_T]$ {Última janela conhecida}
for $h = 1$ **to** H **do**
    $\hat{y}'_{T+h} \leftarrow f_\theta(\mathbf{s})$
    Atualizar janela: $\mathbf{s} \leftarrow [s_2, s_3, \ldots, s_L, \hat{y}'_{T+h}]$
end for
Desnormalizar $\hat{Y}'$ para obter $\hat{Y}$
return $\hat{Y}$

4 Experimentos e Análise de Resultados

4.1 Análise Exploratória da Série Temporal

4.1.1 Histórico de Focos de Incêndio

A Figura 10 apresenta a série temporal mensal do número de focos de incêndio registrados no Pantanal brasileiro, no período compreendido entre 1998 e janeiro de 2026. Observa-se um comportamento marcadamente não estacionário, caracterizado por picos acentuados em anos e meses específicos, com intensificação especialmente a partir da década de 2000.

Destacam-se eventos extremos registrados nos anos de 2005, 2007, 2020 e 2024, os quais evidenciam a ocorrência de episódios severos de queimadas. Esses picos reforçam a complexidade do fenômeno, marcada por elevada variabilidade interanual e sazonal, e indicam a necessidade de modelos capazes de lidar com distribuições assimétricas, alta dispersão e presença de valores extremos.

Adicionalmente, observa-se que os períodos de baixa atividade apresentam valores próximos de zero, o que impõe desafios metodológicos adicionais, especialmente para métricas baseadas em erro relativo, como o MAPE (Mean Absolute Percentage Error). Esse aspecto será discutido de forma mais detalhada nas seções subsequentes, no contexto da avaliação dos modelos preditivos.

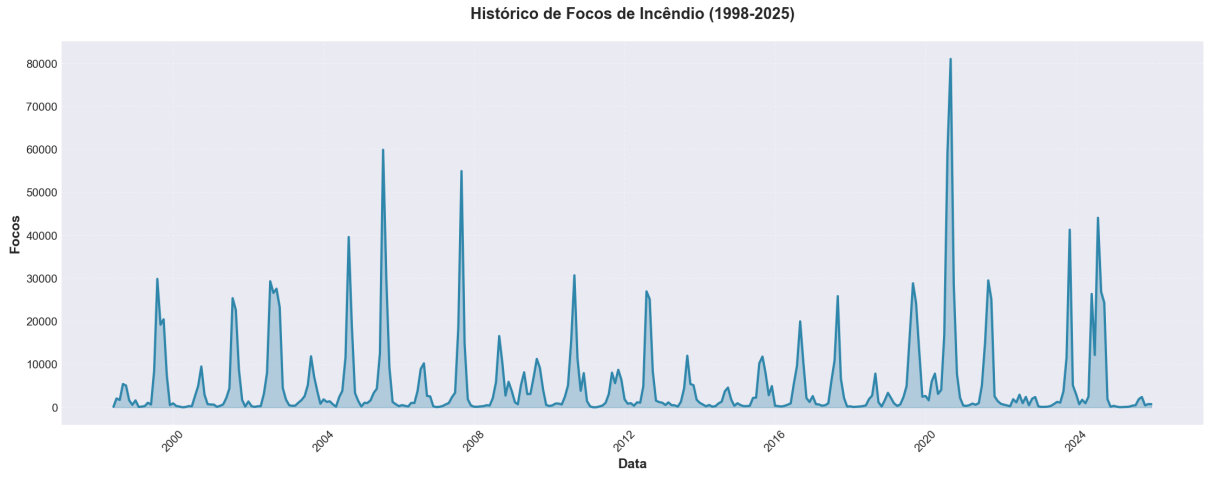

Figura 10 – Histórico de Focos de Incêndio (1998–2025).

4.1.2 Padrões Sazonais e Distribuição Temporal

A Figura 11 apresenta uma análise aprofundada da estrutura sazonal e da distribuição estatística dos focos de incêndio ao longo do período estudado. Diferentemente da Figura 17, que evidenciou a evolução temporal contínua da série histórica, esta figura decompõe o comportamento dos dados sob perspectivas complementares, permitindo uma compreensão mais detalhada da dinâmica do fenômeno.

- a média mensal de focos de incêndio, evidenciando padrões sazonais recorrentes;
- a distribuição estatística por mês, permitindo observar dispersão, assimetria e presença de valores extremos;
- o total anual de focos, destacando variações interanuais e anos críticos;
- e um mapa de calor ano-mês, que sintetiza visualmente a intensidade dos focos ao longo do tempo, facilitando a identificação de ciclos e períodos de maior concentração de ocorrências.

Essa abordagem multifacetada permite identificar tanto a sazonalidade intra-anual, associada ao regime climático do Pantanal, quanto a variabilidade interanual, marcada por episódios extremos de queimadas. Essa análise é fundamental, pois fundamenta diretamente as escolhas realizadas na etapa de engenharia de atributos do modelo XGBoost.

4.1.2.1 Padrão Sazonal Mensal(Média)

O primeiro painel da Figura 11 apresenta a média mensal de focos de incêndio, considerando todo o período analisado (1998–2026). Observa-se um padrão sazonal claramente definido, evidenciando a recorrência de determinados comportamentos ao longo dos anos.

Os meses de junho e julho já apresentam crescimento progressivo no número médio de focos, com intensificação acentuada até atingir o pico nos meses de agosto e setembro. Após esse período, verifica-se uma redução gradual a partir de outubro, acompanhando a transição para condições climáticas mais úmidas.

Esse comportamento confirma a presença de sazonalidade estrutural na série temporal, indicando que o risco de incêndio não se distribui de forma homogênea ao longo do ano, mas apresenta concentração sistemática durante o período seco. Tal característica reforça a importância de modelos preditivos capazes de capturar padrões sazonais recorrentes na dinâmica dos focos de incêndio.

Essa evidência justificou a inclusão, no modelo preditivo, de variáveis relacionadas ao calendário, como:

- número do mês (Mes_num),
- seno e cosseno do mês (Mes_Sin, Mes_Cos),

- ❏ código da estação do ano (Estacao_cod),
- ❏ média histórica do mês (Mes_Media_Historica).

Assim demonstrando que o painel não é apenas descritivo - ele fundamenta matematicamente a necessidade de capturar sazonalidade no modelo.

4.1.2.2 Distribuição Estatística por Mês (Boxplot)

O segundo painel apresenta a distribuição dos valores mensais por meio de boxplots, permitindo analisar não apenas as médias, mas também a variabilidade, a assimetria e a presença de valores extremos ao longo dos anos.

Observa-se que os meses de agosto e setembro, além de registrarem as maiores médias, também exibem maior dispersão interquartil e maior ocorrência de outliers. Esse comportamento indica que, embora exista um padrão sazonal bem definido, a intensidade dos incêndios varia significativamente entre os anos.

Em determinados períodos, registram-se valores excepcionalmente elevados nesses meses, caracterizando eventos extremos de queimadas. Esses episódios reforçam a natureza altamente variável do fenômeno e evidenciam a necessidade de modelos preditivos capazes de lidar com alta variabilidade e ocorrência de extremos na série temporal.

Essa variabilidade reforça a necessidade de incluir no modelo de variáveis que capturem volatilidade e comportamento anômalo, como:

- ❏ desvio padrão (Std_3, Std_6, Std_12),
- ❏ percentil dos últimos 12 meses (Percentil_12m),
- ❏ Z-score 12 meses (ZScore_12m).

Assim, garantindo que o modelo não aprenda apenas médias históricas, mas também identificar quando o comportamento atual está fora do padrão.

4.1.2.3 Total Anual de Focos

O terceiro painel apresenta o total anual de focos de incêndios. Observa-se forte variabilidade interanual, com destaque para anos como 2005 e 2020, que registraram volumes significamente superiores aos demais.

Isso indica que a série não apresenta crescimento linear simples, mas sim ciclos e eventos críticos concentrados em determinados anos. Essa irregularidade interanual justifica o uso de variáveis que capturam dependência temporal de longo prazo, como:

- ❏ Lag_12 (valor do mesmo mês anterior),
- ❏ Diff_12 (diferença em relação ao mesmo período do ano anterior),

- Desvio_Media_Historica

Esse painel evidencia que o comportamento da série depende fortemente da memória anual, o que é coerente com fenômenos ambientais influenciados por ciclos climáticos.

4.1.2.4 Heatmap Ano x Mês

O quarto painel apresenta um mapa de calor com a intensidade de focos distribuída por ano e mês.

As cores mais intensas concentram-se nos meses de agosto e setembro, reforçando o padrão sazonal já identificado. Além disso, determinados anos apresentam coloração muito mais intensa que os demais, evidenciando eventos extremos concentrados temporalmente.

O heatmap permite visualizar simultaneamente:

- a estrutura intra-anual (sazonalidade mensal),
- a estrutura interanual (variação entre anos).

Essa vizualização confirma o comportamento da série é dependente tanto do mês quanto do contexto anual, reforçando a necessidade de variáveis que combinem informação recente e histórica.

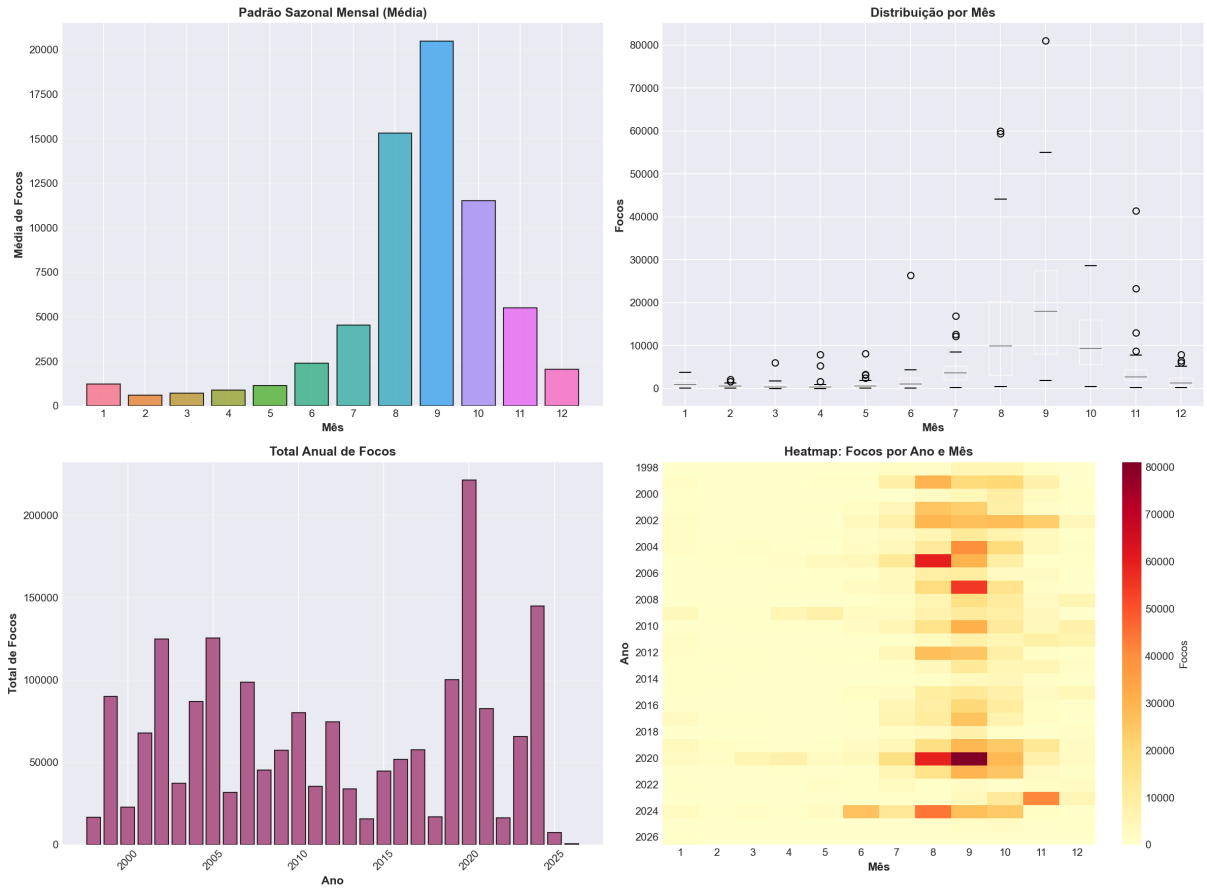

Figura 11 – Padrões Sazonais e Distribuição Temporal.

4.2 Engenharia de Atributos e Importância de Variáveis

4.2.1 Importância das Features

A Figura 12 apresenta o ranking das 20 variáveis mais importantes, segundo o critério de ganho (gain) do modelo XGBoost. Observa-se que a variável Desvio_Media_Historica concentra aproximadamente metade da importância total do modelo, indicando que a diferença entre o valor recente e a média histórica do respectivo mês desempenha papel central no processo de previsão.

Entre as demais variáveis de destaque, incluem-se:

- `Diff_12`, que captura a variação em relação ao mesmo período do ano anterior;
- `ZScore_12m`, que representa o desvio padronizado do comportamento recente em relação à média histórica;
- `Std_3` e `Diff_1`, associadas à volatilidade e às variações de curto prazo.

Esses resultados indicam que o modelo aprende predominantemente a partir de anomalias em relação ao padrão histórico, e não apenas de valores absolutos da série. Tal evidência reforça a adequação da estratégia de engenharia de atributos adotada no Capítulo 3, especialmente no que se refere à incorporação de variáveis derivadas que capturam tendência, sazonalidade e volatilidade.

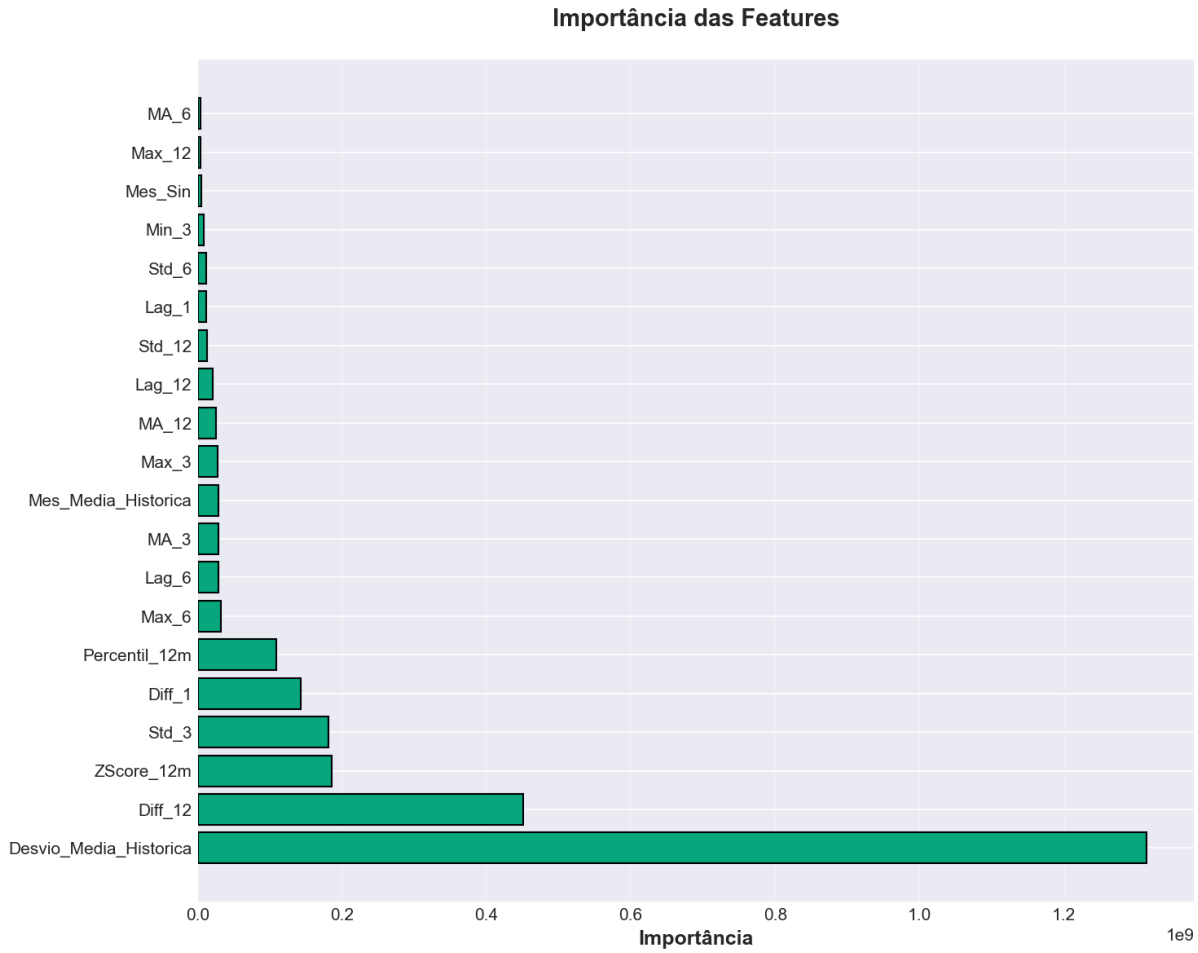

Figura 12 – Padrões Sazonais e Distribuição Temporal.

4.3 Avaliação Preditiva do Modelo

4.3.1 Previsões versus Valores Reais

A Figura 13 apresenta a relação entre os valores previstos e os valores observados no conjunto de teste. Verifica-se uma forte concentração dos pontos ao redor da linha de identidade (45^{o}), indicando elevada correlação entre previsões e valores reais.

O coeficiente de determinação ($R^2 = 0,9372$) confirma que o modelo é capaz de explicar grande parte da variabilidade observada na série, mesmo em um contexto caracterizado por alta dispersão e presença de eventos extremos.

Entretanto, observa-se maior dispersão nos valores mais elevados de focos de incêndio, sugerindo que o modelo tende a suavizar picos extremamente altos. Esse comportamento é esperado em métodos baseados em árvores, especialmente quando treinados com séries temporais relativamente curtas, nas quais eventos extremos representam uma pequena fração do conjunto de treinamento.

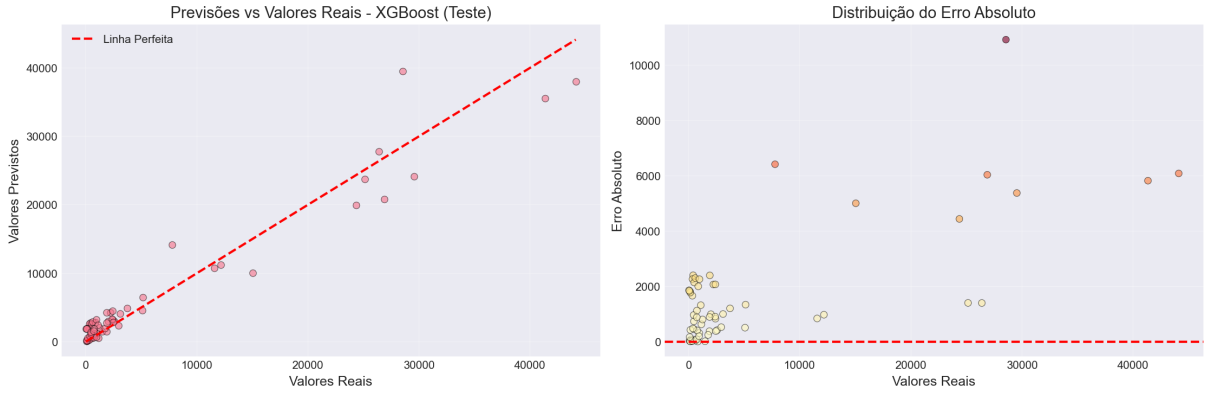

Figura 13 – Padrões Sazonais e Distribuição Temporal.

4.3.2 Análise dos Resíduos

A Figura 14 apresenta diferentes perspectivas da análise dos resíduos do modelo no conjunto de teste. Observa-se que a média dos resíduos é próxima de zero, indicando a ausência de viés sistemático significativo nas previsões.

Entretanto, a distribuição dos resíduos apresenta caudas alongadas e elevada curtose, refletindo a dificuldade do modelo em capturar com exatidão eventos extremos raros. O gráfico Q–Q reforça essa interpretação ao evidenciar desvios em relação à normalidade, especialmente nas extremidades da distribuição.

Apesar dessas características, a análise temporal dos resíduos não revela padrões persistentes ou autocorrelação evidente, sugerindo que o modelo é capaz de capturar adequadamente a estrutura temporal dominante da série, incluindo tendência e sazonalidade.

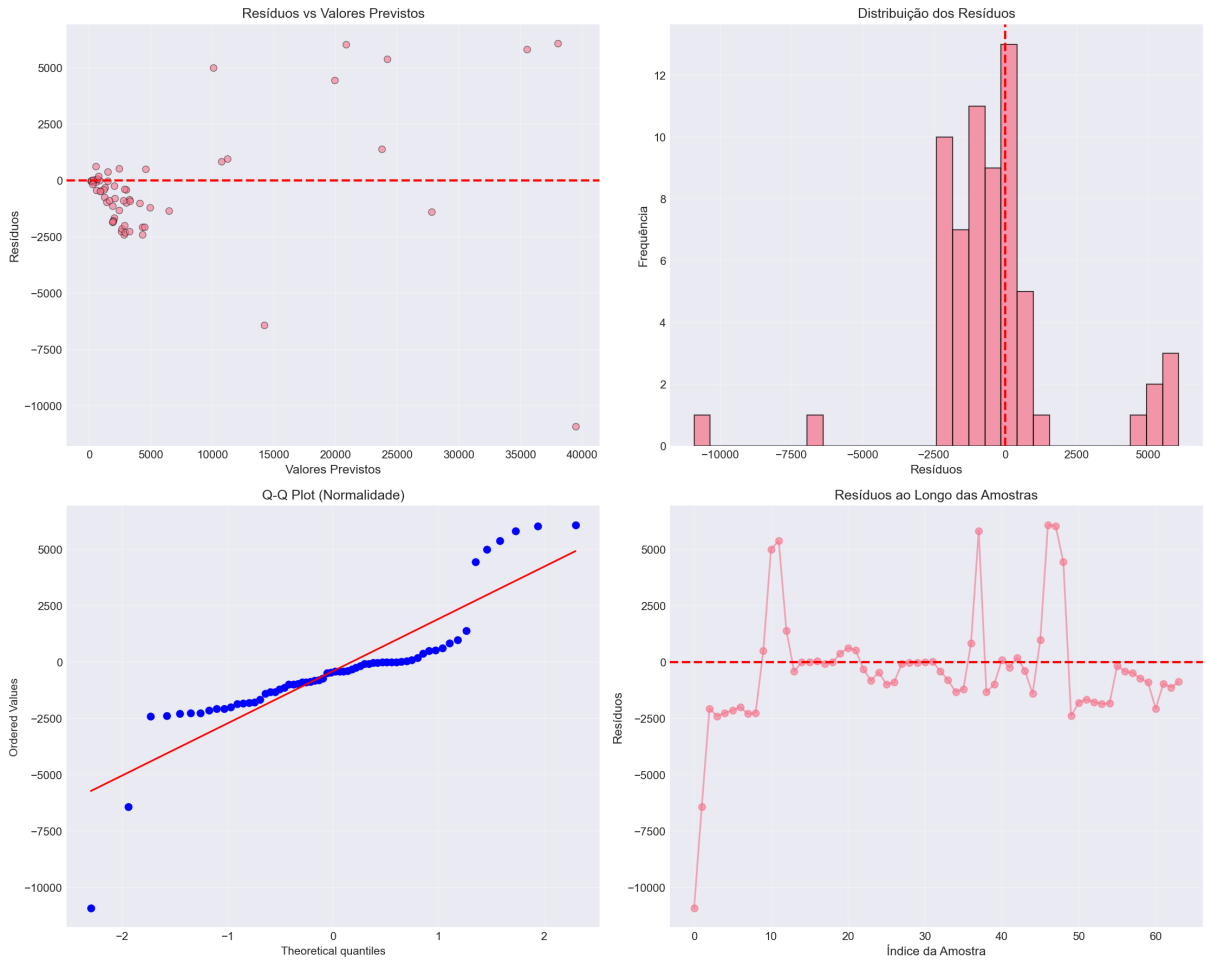

Figura 14 – Análise dos Resíduos.

4.4 Avaliação Temporal das Previsões

4.4.1 Previsões ao Longo do Tempo

A Figura 15 apresenta a comparação entre os valores observados e os valores previstos no conjunto de teste. Observa-se que o modelo é capaz de acompanhar adequadamente a dinâmica sazonal da série, reproduzindo os ciclos recorrentes de crescimento e redução no número de focos ao longo dos anos avaliados.

As maiores discrepâncias concentram-se nos meses associados a picos abruptos de incêndio, caracterizados por variações intensas em curto intervalo de tempo. Esse comportamento pode ser atribuído tanto às limitações inerentes do modelo na antecipação de eventos extremos raros quanto à ausência de variáveis exógenas explicativas, tais como condições meteorológicas (temperatura, umidade, vento) e fatores antrópicos relacionados ao uso e ocupação do solo.

Ainda assim, a capacidade do modelo de reproduzir a estrutura sazonal dominante

da série reforça sua adequação como ferramenta de apoio à previsão do comportamento médio dos focos de incêndio no Pantanal.

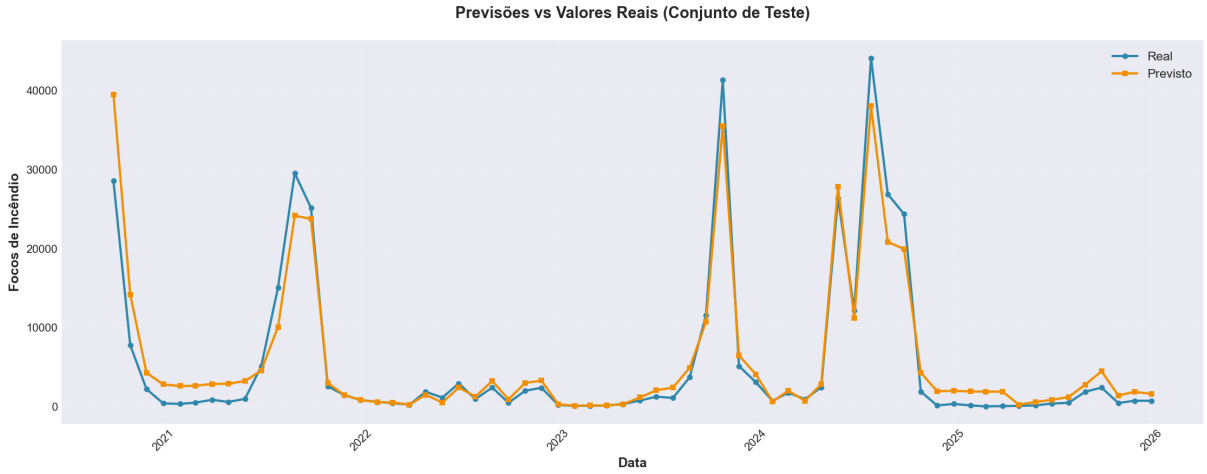

Figura 15 – Previsões versus Valores Reais.

4.5 Dashboard Integrado de Resultados

A Figura 16 consolida os principais resultados obtidos no estudo em um painel único, integrando histórico da série, padrões sazonais, relação entre valores reais e previstos, importância das variáveis e métricas de desempenho.

As métricas obtidas no conjunto de teste indicam:

1. RMSE ≈ 2570 $focos$
2. MAE ≈ 1600 $focos$
3. $R^2 \approx 0{,}94$

Embora o valor do MAPE seja elevado, esse resultado deve ser interpretado com cautela, uma vez que a série apresenta meses com valores muito baixos de focos, inflacionando artificialmente o erro percentual.

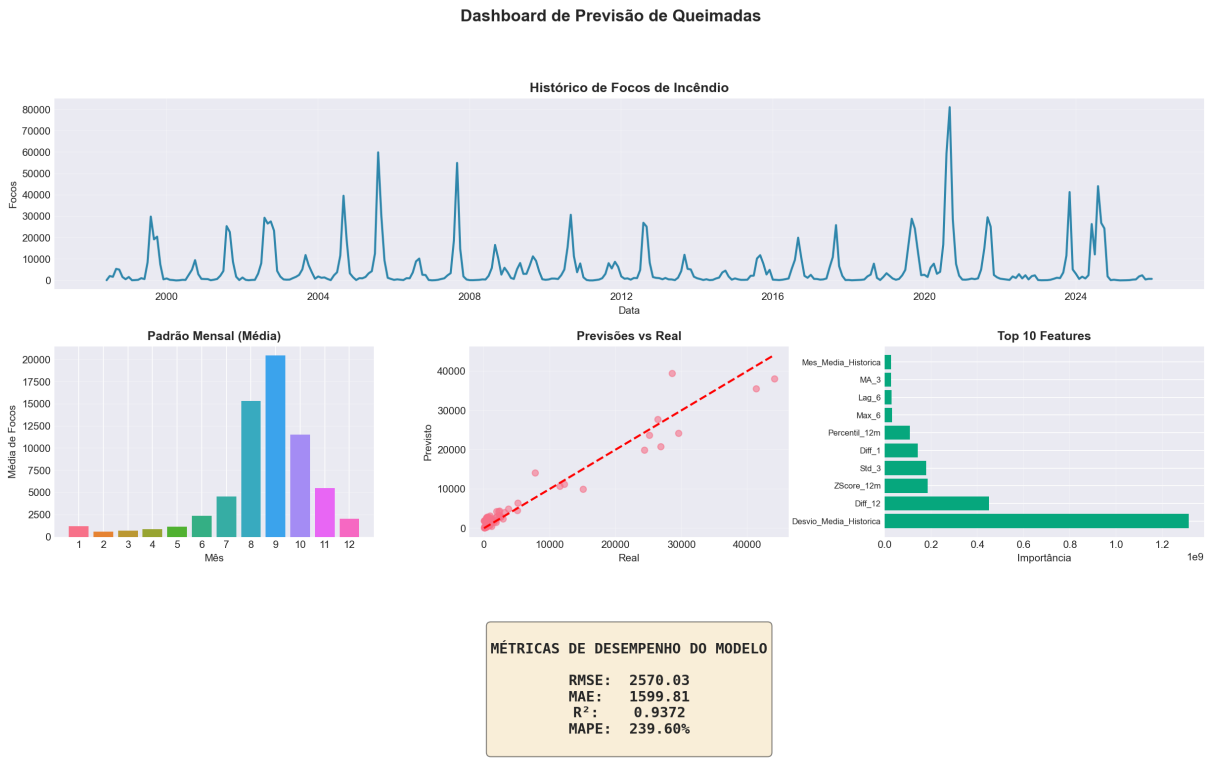

Figura 16 – Dashboard de Previsão de Queimadas.

4.6 Discussão Geral dos Resultados

De forma geral, os resultados indicam que o modelo XGBoost apresentou desempenho satisfatório na previsão do número de focos de incêndio no Pantanal brasileiro, destacando-se especialmente na captura de padrões sazonais recorrentes e de tendências de médio prazo presentes na série temporal.

As limitações observadas estão associadas principalmente a três fatores centrais:

1. O tamanho relativamente reduzido da base histórica, considerando a escala temporal mensal e a ocorrência esparsa de eventos extremos;

2. A ausência de variáveis exógenas climáticas e ambientais, como temperatura, umidade relativa, vento e indicadores de uso do solo, que poderiam contribuir para maior poder explicativo;

3. A natureza altamente irregular e episódica dos eventos extremos, cuja ocorrência depende de múltiplos fatores não totalmente representados no histórico da série.

Apesar dessas restrições, o modelo demonstrou robustez estatística, boa capacidade preditiva e elevada interpretabilidade, característica inerente a métodos baseados em árvores. Tais atributos reforçam seu potencial de aplicação prática, especialmente como

ferramenta de apoio ao planejamento estratégico e à tomada de decisão por órgãos de monitoramento ambiental.

4.6.1 Avaliação Preditiva do modelo LSTM

4.6.1.1 Previsões versus valores reais

A Figura 17 apresenta a relação entre os valores previstos pelo modelo LSTM e os valores observados no conjunto de teste. Idealmente, os pontos deveriam se concentrar próximos à linha de identidade (45°), indicando previsões consistentes ao longo de toda a faixa de valores.

Entretanto, observa-se um comportamento distinto do esperado:

Primeiramente, há elevada dispersão dos pontos, especialmente à medida que os valores reais aumentam. Isso indica dificuldade do modelo em manter precisão quando confrontado com meses de maior intensidade de focos.

Em segundo lugar, identifica-se uma tendência do modelo a "comprimir" as previsões para valores elevados. Em meses críticos, caracterizados por picos extremos, o LSTM frequentemente prevê valores substancialmente inferiores aos observados, evidenciando um comportamento sistemático de subestimação de extremos.

Além disso, a maior concentração da nuvem de pontos em valores baixos e médios sugere que o modelo apresenta melhor desempenho no regime considerado "regular" da série — isto é, meses com poucos focos —, mas encontra maior dificuldade quando ocorrem eventos raros e abruptos.

Esse comportamento é refletido diretamente nas métricas de desempenho obtidas no conjunto de teste:

1. RMSE ≈ 9607 e MAE ≈ 6452, indicando erros absolutos expressivos, especialmente relevantes em uma série caracterizada por picos elevados;

2. $R^2 \approx 0{,}1218$, o que demonstra que o modelo explica apenas uma parcela limitada da variabilidade observada no conjunto de teste, sugerindo que a dinâmica completa — principalmente na região extrema — não é adequadamente capturada;

3. RMSE ≈ 953, métrica que se mostra pouco informativa neste contexto, pois a série apresenta meses com valores muito próximos de zero. Nessas situações, mesmo erros absolutos moderados geram erros percentuais extremamente elevados, inflacionando artificialmente o indicador.

De forma geral, os resultados indicam que o modelo LSTM é capaz de aprender um padrão médio da série, porém não demonstra confiabilidade para reproduzir a amplitude real dos picos extremos, que representam justamente os eventos de maior interesse para fins de monitoramento e planejamento estratégico.

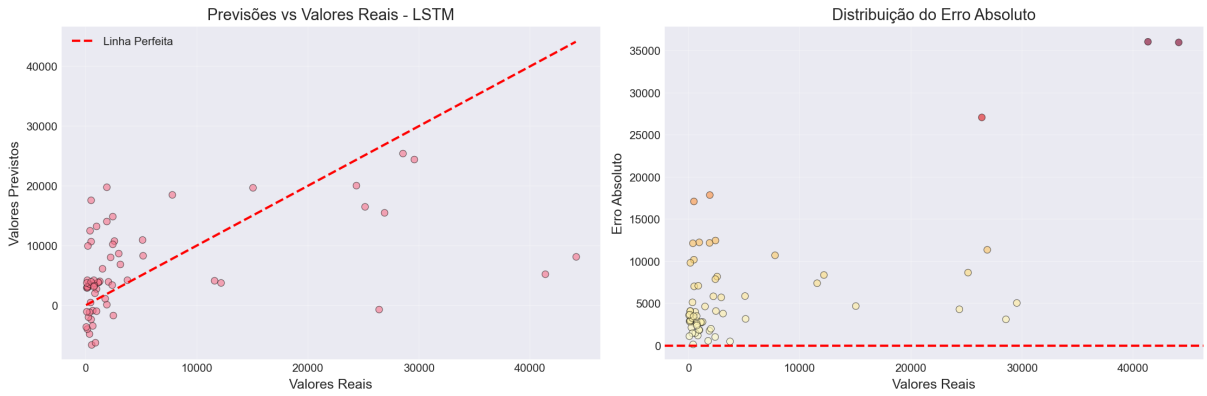

Figura 17 – Previsões versus Valores Reais.

4.6.2 Análise dos Resíduos

A Figura 18 apresenta quatro perspectivas complementares da análise de resíduos (resíduo = valor real - valor previsto), permitindo compreender de forma mais aprofundada as limitações do modelo LSTM neste cenário.

1. **Resíduos vs valores previstos**

   Observam-se indícios de heterocedasticidade, isto é, a variância dos erros não permanece constante ao longo da faixa de valores previstos. Em modelos bem ajustados, espera-se uma dispersão aproximadamente aleatória dos resíduos em torno de zero. Contudo, verifica-se um padrão estruturado, com aumento da magnitude do erro em determinados regimes. Esse comportamento sugere que o modelo não está igualmente calibrado em toda distribuição, apresentando desempenho desigual conforme o nível da variável alvo.

2. **Distribuição dos resíduos**

   A distribuição não é simétrica/normal. Isso aparece também nas estatísticas:

   - ❏ Skewness $\approx 1,81$ (assimetria positiva)
   - ❏ Kurtosis $\approx 5,61$ (caudas pesadas)

   Ou seja, o erro tem caudas longas, típico de séries com outliers e choques (picos de queimadas) que o modelo não consegue prever com precisão

3. **Q-Q plot**

   Os desvios nas extremidades reforçam que os resíduos não seguem normalidade - novamente compatível com eventos extremos raros e com a dificuldade do modelo em capturar a magnitude correta dos picos.

4. **Resíduos ao longo das amostras**

   Há trechos com erros consecutivos do mesmo sinal e "rajadas"de erro em períodos específicos. Isso sugere mudanças de regime (anos/momentos em que o padrão muda) e/ou sazonalidade com amplitude variável, algo difícil para um LSTM simples e univariado capturar usando apenas o histórico imediato

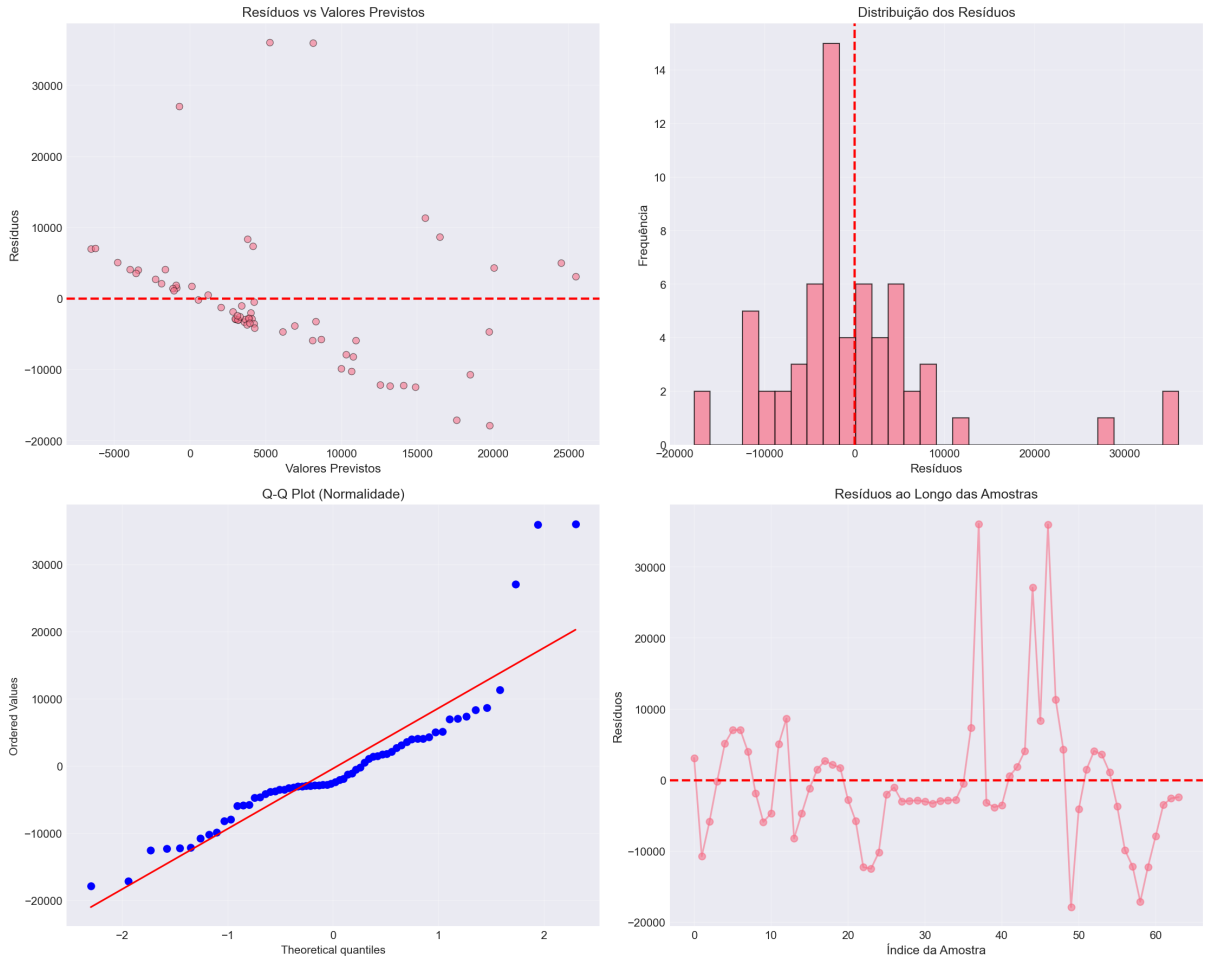

Figura 18 – Análise de Resíduos.

4.6.3 Avaliação Temporal das previsões

A Figura 19, que apresenta a comparação entre valores reais previstos ao longo do tempo, constitui uma das análises mais elucidativas do desempenho do modelo, pois permite visualizar diretamente como e quando os erros ocorrem.

Observa-se que o modelo LSTM é capaz de acompanhar a sazonalidade geral da série, reproduzindo adequadamente os ciclos recorrentes de elevação ao longo dos anos. Esse comportamento indica que a rede conseguiu aprender a estrutura básica do ciclo anual, capturando o padrão sistemático associado ao período seco.

Entretanto, nos momentos caracterizados por picos agudos de incêndio, o modelo tende a suavizar a magnitude das previsões, frequentemente atrasando ou antecipando parcialmente o pico e reduzindo sua amplitude. Em termos práticos, isso significa que o modelo aprende "quando a série costuma subir", mas não consegue reproduzir com precisão "o quanto ela sobe"nos anos críticos.

Esse comportamento é coerente com a própria natureza do fenômeno analisado. Incêndios extremos possuem caráter episódico e multifatorial, sendo fortemente influenciados por variáveis exógenas como condições climáticas(temperatura, umidade, vento), períodos de seca prolongada, uso do solo e ação antrópica. Como tais eventos não ocorrem com frequência suficiente para formar padrões recorrentes robustos apenas a partir da série histórica de focos, o modelo univariado apresenta limitações naturais em sua capacidade de antecipar esses extremos.

De modo geral, a análise temporal reforça que o LSTM captura adequadamente a dinâmica estrutural média da série, mas demonstra fragilidade na modelagem da variabilidade extrema e dos choques abruptos.

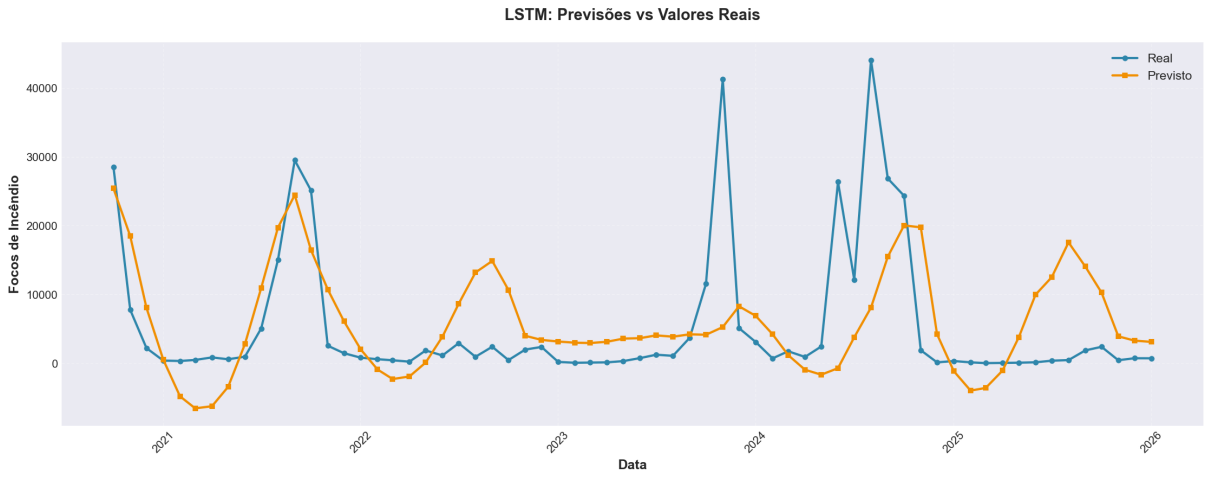

Figura 19 – LSTM: Previsões versus Valores Reais.

4.6.4 Discussão Geral dos Resultados

Os resultados obtidos indicam que o modelo LSTM apresentou funcionamento adequado dentro das condições metodológicas estabelecidas neste estudo. O processo de treinamento convergiu de forma estável, com estabilização das curvas de loss e val_loss, além da atuação do mecanismo de early stopping, evidenciando ajuste consistente aos dados disponíveis.

Do ponto de vista preditivo, o modelo demonstrou capacidade de capturar o padrão médio e a sazonalidade anual da série temporal, reproduzindo os ciclos recorrentes de crescimento e queda ao longo do tempo. Ademais, a média dos resíduos próxima de

zero sugere ausência de viés sistemático relevante, indicando que o modelo não superestima ou subestima de maneira persistente os valores previstos. As limitações observadas concentram-se, portanto, na calibração dos extremos e na variabilidade dos erros, e não propriamente na presença de viés estrutural.

Entretanto, alguns fatores técnicos ajudam a explicar por que o desempenho global não foi superior. Embora a série histórica abranja um período longo, a quantidade efetiva de amostras disponíveis para treinamento — após a construção das sequências temporais — é relativamente reduzida para um modelo neural com elevado número de parâmetros. Essa relação entre complexidade do modelo e volume de dados pode comprometer a capacidade de generalização, especialmente em uma série caracterizada por ruído e caudas pesadas.

Além disso, a modelagem adotada foi univariada, utilizando exclusivamente o histórico de focos de incêndio como variável de entrada. Considerando que eventos extremos de queimadas dependem fortemente de fatores exógenos — como precipitação, temperatura, umidade relativa, vento, condições de seca e uso do solo —, a ausência dessas variáveis limita a capacidade do modelo de antecipar choques abruptos.

A própria distribuição dos dados também impõe desafios adicionais. A elevada assimetria e a discrepância entre valores medianos e máximos indicam presença de extremos raros e de grande magnitude. Nesse contexto, o modelo tende a aprender com maior precisão o regime considerado "normal", apresentando subestimação dos picos mais severos. A normalização por MinMax, embora adequada em termos gerais, pode intensificar esse efeito ao comprimir a maior parte dos valores em uma faixa reduzida quando existem picos muito elevados.

Por fim, a definição de uma janela temporal fixa de 12 meses, embora coerente para capturar sazonalidade anual, pode não ser suficiente para modelar dependências interanuais mais longas, como períodos prolongados de seca ou mudanças estruturais associadas a fatores climáticos e antrópicos.

Em síntese, o modelo LSTM não apresentou falha metodológica, mas sim limitações inerentes ao contexto dos dados e à estratégia de modelagem adotada. A rede foi capaz de reproduzir adequadamente o comportamento sazonal médio da série; contudo, seu desempenho é restringido pela combinação entre reduzido volume efetivo de treinamento, ausência de variáveis exógenas e ocorrência de eventos extremos raros e altamente assimétricos.

4.6.5 Comparação entre os modelos LSTM e XGBoost

Quando comparado ao modelo XGBoost apresentado neste capítulo, o LSTM apresentou um desempenho inferior no conjunto de teste. Enquanto o LSTM obteve coeficiente de determinação aproximado de $R^2 \approx 0{,}1218$, o XGBoost alcançou valor substancialmente superior, indicando maior capacidade explicativa da variabilidade observada na série.

Essa diferença de desempenho é consistente com as características metodológicas de cada abordagem. O modelo XGBoost foi alimentado por um conjunto estruturado de variáveis derivadas por meio de engenharia de atributos, incluindo defasagens temporais (lags), médias móveis, estatísticas de volatilidade e transformações sazonais (como seno e cosseno do mês). Essa estratégia torna explícitas informações sobre sazonalidade, tendência e anomalias, facilitando o aprendizado em bases tabulares de dimensão reduzida.

Por outro lado, o LSTM depende de maior volume de dados e/ou da presença de múltiplas variáveis explicativas para construir representações internas robustas. Em um cenário univariado e caracterizado por eventos extremos raros, o modelo tende a capturar adequadamente o comportamento médio da série, porém apresenta tendência à suavização dos picos, reduzindo sua capacidade de explicar a variabilidade total dos dados.

Dessa forma, no contexto deste estudo, o LSTM pode ser interpretado como um baseline neural consistente, demonstrando que há sinal temporal e sazonalidade aprendíveis apenas a partir da série histórica. Contudo, na ausência de enriquecimento das entradas e diante da limitada quantidade de exemplos extremos, o modelo não supera a abordagem baseada em árvores com engenharia de atributos explícita.

Essa comparação reforça a importância da adequação entre método e estrutura dos dados, evidenciando que, para séries temporais univariadas com amostras limitadas e alta assimetria, modelos baseados em árvores combinados com engenharia de atributos podem apresentar vantagem prática significativa.

5 Conclusão

O presente estudo teve como objetivo avaliar o desempenho de modelos de aprendizado de máquina aplicados à previsão de séries temporais de focos de incêndio no Pantanal brasileiro, utilizando dados mensais compreendidos entre 1998 e janeiro de 2026. Para tal, foram implementadas e comparadas duas abordagens metodologicamente distintas: um modelo baseado em árvores de decisão com técnica de gradient boosting (XGBoost) e uma rede neural recorrente do tipo Long Short-Term Memory (LSTM).

A análise exploratória evidenciou que a série apresenta forte sazonalidade anual, elevada variabilidade interanual e ocorrência de eventos extremos episódicos, caracterizados por distribuição altamente assimétrica e presença de picos abruptos. Essas características impõem desafios relevantes à modelagem preditiva, especialmente quando se dispõe de base univariada.

Os resultados indicaram que o modelo XGBoost apresentou desempenho superior no conjunto de teste, com menores valores de erro médio absoluto (MAE) e raiz do erro quadrático médio (RMSE), além de maior estabilidade preditiva. Esse desempenho pode ser atribuído à sua robustez frente a não linearidades, à presença de mecanismos de regularização e à eficácia da engenharia de atributos temporais adotada, que tornou explícitas informações relacionadas à sazonalidade e às variações interanuais.

Por sua vez, o modelo LSTM demonstrou capacidade consistente de capturar o padrão sazonal médio da série e apresentou convergência estável durante o treinamento. Entretanto, seu desempenho foi limitado por fatores estruturais do conjunto de dados. A principal dificuldade enfrentada neste estudo foi a disponibilidade restrita de variáveis explicativas. A base utilizada contempla exclusivamente o número de focos ativos de incêndio, configurando uma modelagem univariada.

Modelos de deep learning, como a LSTM, tendem a apresentar melhor desempenho quando alimentados com múltiplas variáveis que representem diferentes dimensões do fenômeno estudado. No contexto de incêndios florestais, variáveis climáticas e ambientais — como precipitação, temperatura do ar, umidade relativa, velocidade do vento e índices de vegetação — exercem influência direta sobre a ocorrência e intensidade dos focos. A ausência dessas informações limita a capacidade da rede de capturar os fatores que desencadeiam eventos extremos, restringindo o aprendizado a padrões temporais internos da própria série.

Além disso, o volume efetivo de amostras para treinamento mostrou-se relativamente reduzido para uma arquitetura neural com milhares de parâmetros, o que também impacta a capacidade de generalização do modelo. Dessa forma, a limitação observada não decorre de falha metodológica, mas da natureza dos dados disponíveis e da complexidade intrínseca do fenômeno modelado.

Do ponto de vista metodológico, a comparação entre um modelo ensemble baseado em árvores e uma rede neural recorrente permitiu avaliar duas abordagens distintas de tratamento da dependência temporal: a modelagem explícita via engenharia de atributos e a modelagem implícita por meio de memória sequencial. Para o conjunto de dados analisado, os resultados sugerem que abordagens baseadas em boosting apresentam melhor relação entre desempenho preditivo e custo computacional.

De maneira geral, conclui-se que, embora a previsão precisa de eventos extremos permaneça um desafio significativo, os modelos desenvolvidos demonstram potencial como ferramentas de apoio ao monitoramento ambiental e ao planejamento estratégico. A incorporação de variáveis exógenas e a ampliação da base de dados representam caminhos promissores para aprimorar a capacidade preditiva em estudos futuros.

Adicionalmente, reconhecem-se como limitações deste estudo a utilização de modelagem univariada, a ausência de variáveis climáticas e ambientais complementares, a não incorporação explícita de fatores socioeconômicos relacionados ao uso do solo e a possível influência de mudanças climáticas de longo prazo não modeladas diretamente. Como desdobramento da pesquisa, recomenda-se a ampliação do modelo para abordagem multivariada, incorporando dados meteorológicos e índices ambientais, o teste de arquiteturas híbridas (como CNN-LSTM), a utilização de funções de perda mais robustas a outliers e a adoção de modelos probabilísticos que permitam estimar incertezas associadas às previsões. Tais avanços poderão contribuir para maior precisão na antecipação de períodos críticos e para o fortalecimento das estratégias de prevenção e mitigação de incêndios no Pantanal brasileiro.

Referências

ABBES, A. B.; MAGAGI, R.; GOITA, K. Soil moisture estimation from smap observations using Long Short-Term Memory (LSTM). 2019.

AHSAN, M. M. et al. Effect of data scaling methods on machine learning algorithms and model performance. Technologies, v. 9, n. 3, 2021. ISSN 2227-7080.

AKTER, R.; LEE, J. M.; KIM, D. S. Analysis and prediction of hourly energy consumption based on long short-term memory neural network. International Conference on Information Networking (ICOIN), p. 732–734, 2021.

ALPAYDIN, E. Introduction to Machine Learning. [S.l.]: Adaptive Computation and Machine Learning MIT Press, 2014.

ALVARES, C. A. et al. Perigo de incêndio florestal: aplicação da fórmula de monte alegre e avaliação do histórico para Piracicaba, SP. Scientia Forestalis, v. 42, n. 104, p. 521–532, 2014.

AMBARWARI, A.; ADRIAN, Q. J.; HERDIYENI, Y. Analysis of the effect of data scaling on the performance of the machine learning algorithm for plant identification. Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi), v. 4, n. 1, p. 117–122, 2020.

ANGSTROM, A. Riskerna for skogsbrand och deras beroende av vader och klimat ["The risks for forest fires and their relation to weather and climate"]. [S.l.]: Svenska Skogsvirdsforeningens Tidskrift, 1942. v. 40. 323–343 p.

BATISTA, G.; BAZZAN, A. L. C.; MONARD, M. C. Balancing training data for automated annotation of keywords: a case study. In: WOB. [S.l.: s.n.], 2003. p. 10–18.

BATISTA, G.; PRATI, R. C.; MONARD, M. C. A study of the behavior of several methods for balancing machine learning training data. ACM Sigkdd Explorations Newsletter, v. 6, n. 1, p. 20–29, 2004.

BRYSON, A. E.; HO, Y. C. Applied Optimal Control. [S.l.: s.n.], 1969.

BUDA, M.; MAKI, A.; MAZUROWSKI, M. A. A systematic study of the class imbalance problem in convolutional neural networks. Neural Networks, v. 106, p. 249–259, 2018. ISSN 0893-6080.

BURKOV, A. The Hundred-Page Machine Learning Book. [S.l.]: Andriy Burkov, 2019. ISBN 9781999579517.

CERRI, R. Técnicas de Classificação Hierárquica Multirrótulo. Dissertação

(Dissertação de Mestrado) — Instituto de Ciências Matemáticas e de Computação,

Universidade de São Paulo- ICMC USP, 2010.

CHAWLA, N. V. et al. SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, v. 16, p. 321–357, 2002.

CHEN, T.; GUESTRIN, C. Xgboost: A scalable tree boosting system. ACM SIGKDD

International Conference on Knowledge Discovery and Data Mining, p.

785––794, 2016.

CORTES, C.; VAPNIK, V. Support-vector networks. Machine Learning, Springer, v. 20, n. 3, p. 273–297, 1995.

CRISTIANINI, N.; SHAWE-TAYLOR, J. An Introduction to Support Vector Machines and Other Kernel-based Learning Methods. [S.l.]: Cambridge University Press, 2000.

CUNNINGHAM, P.; DELANY, S. J. k-Nearest Neighbour Classifiers: 2nd Edition (with Python examples). Computer Vision and Pattern Recognition (CoRR), 2020.

DIETTERICH, T. G. Ensemble methods in machine learning. In: Multiple Classifier Systems. [S.l.]: Springer Berlin Heidelberg, 2000. p. 1–15.

DIMRI, T.; AHMAD, S.; SHARIF, M. Time series analysis of climate variables using seasonal arima approach. Journal of Earth System Science, v. 129, n. 1, p. 149, 2020.

GAO, C.; LIN, H.; HU, H. Forest-fire-risk prediction based on random forest and backpropagation neural network of heihe area in heilongjiang province, china. Forests 2023, v. 14, n. 2, 2023.

HAN, H.; WANG, W.; MAO, B. Borderline-smote: A new over-sampling method in imbalanced data sets learning. In: . [S.l.: s.n.], 2005. p. 878–887.

HE, H. et al. Adasyn: Adaptive synthetic sampling approach for imbalanced learning.

In: . [S.l.: s.n.], 2008. p. 1322–1328.

HOLLAND, J. H. Adaptation in Natural and Artificial Systems. [S.l.]: University of Michigan Press, 1975.

INPE. Queimadas: monitoramento de focos. 2026. <terrabrasilis.dpi.inpe.br/queimadas/situação-atual/estatistica/estatatistica_estados/> Acesso em: 26 de Janeiro de 2026.

JOHNSON, J.; KHOSHGOFTAAR, T. The effects of data sampling with deep learning and highly imbalanced big data. Information Systems Frontiers, v. 22, 2020.

KE, G. et al. Lightgbm: A highly efficient gradient boosting decision tree. Advances in neural information processing systems, v. 30, p. 3146–3154, 2017.

KOPRINSKA, I.; WU, D.; WANG, Z. Convolutional neural networks for energy time series forecasting. In: . [S.l.: s.n.], 2018. p. 1–8.

KUBAT, M.; MATWIN, S. Addressing the curse of imbalanced training sets: one-sided selection. In: ICML. [S.l.: s.n.], 1997. v. 97, p. 179–186.

LAURIKKALA, J. Improving identification of difficult small classes by balancing class distribution. In: Springer Berlin Heidelberg. [S.l.: s.n.], 2001.

LEMAITRE, G.; NOGUEIRA, F.; ARIDAS, C. K. Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning. 2016.

LIN, X. et al. Forest fire prediction based on long- and short-term time-series network. Forests, v. 14, n. 4, 2023.

MASSMANN, C.; HOLZMANN, H. Analysing goodness of fit measures using a sensitivity based approach. General Assembly Conference Abstracts, pp. 12354, p. 12354–, 04 2012.

MUKADI, P. M.; GONZÁLEZ-GARCÍA, C. Time series analysis of climatic variables in peninsular spain. trends and forecasting models for data between 20th and 21st centuries. Climate, v. 9, n. 7, 2021. ISSN 2225-1154.

NARCISO, M.; SORIANO, B. Saripan- sistema de avaliação de risco de incêndio para o pantanal. Congresso Brasileiro de Agroinformática, 2019.

NESTEROV, V. G. Combustibility of the forest and methods for its determination (in russian). USSR State Industry Press, 1949.

NUNES, J. R. S. et al. Desempenho da fórmula de monte alegre (FMA) e da fórmula de monte alegre alterada (FMA+) no distrito florestal de monte alegre. Revista Floresta, 2010. ISSN 1982-4688.

NUNES, J. R. S.; SOARES, R. V.; BATISTA, A. C. FMA+- Um novo índice de perigo de incêndios florestais para o o estado do Paraná, Brasil. [S.l.]: Revista Floresta, 2006. ISSN 1982-4688.

ONIGEMO, A. E. Avaliação de índices de risco de incêndio em áreas com predominância de gramíneas cespitosas na sub-região da Necholândia, Pantanal, MS. Tese (Doutorado em Ecologia) — Programa de Pós-graduação em Ecologia e Conservação, Universidade Federal do Mato Grosso do Sul, 2007.

P., P. P.; VANITHA, V.; R, R. Wind speed forecasting using long short term memory networks. 2nd International Conference on Intelligent Computing, Instrumentation and Control Technologies (ICICICT), p. 1310–1314, 2019.

PARMEZAN, A. R.; SOUZA, V. M. A.; BATISTA, G. E. A. P. A. Evaluation of statistical and machine learning models for time series prediction: Identifying the state-of-art and the best conditions for the use of each model. Information Sciences, v. 484, p. 302–337, 2019.

POPESCU,M.etal. Multilayer perceptron and neural networks. WSEAS Transactions on Circuits and Systems, v. 8, 2009.

PROKHORENKOVA, L. et al. Catboost: unbiased boosting with categorical features. Conference on Neural Information Processing Systems (NeurIPS), 2018.

QUINLAN, J. R. Induction of decision trees. Mach. Learn., Kluwer Academic Publishers, USA, v. 1, n. 1, p. 81––106, 1986. ISSN 0885-6125.

_____. C4.5: Programs for Machine Learning. [S.l.]: Morgan Kaufmann Publishers Inc., 1993. ISBN 1-55860-238-0.

RAKSHIT, P. et al. Prediction of forest fire using machine learning algorithms: The search for the better algorithm. In: 2021 6th International Conference on Innovative Technology in Intelligent System and Industrial Applications (CITISIA). [S.l.: s.n.], 2021. p. 1–6.

ROSENBLATT, F. The Perceptron, a Perceiving and Recognizing Automaton. [S.l.]: Cornell Aeronautical Laboratory, 1957.

RUBÍ, J. N.; CARVALHO, P. H. de; GONDIM, P. R. Application of machine learning models in the behavioral study of forest fires in the brazilian federal district region. Engineering Applications of Artificial Intelligence, v. 118, p. 105649, 2023.

RUSSELL, S. J.; NORVIG, P. Artificial intelligence: a modern approach. [S.l.]: Pearson, 2010.

SAMPAIO, O. B. Estudo Comparativo de Índice, para Previsão de Incêndios Florestais, na Região de Coronel Fabriciano, Minas Gerais. Dissertação (Dissertação de Mestrado) — Departamento de Engenharia Florestal. Universidade Federal de Vićosa, 1991.

SANTOS, B. Z. et al. A new time series framework for forest fire risk forecasting and classification. In: 2023 International Joint Conference on Neural Networks (IJCNN). [S.l.: s.n.], 2023.

SCHAPIRE, R. E. Explaining ADABoost. In: Empirical inference. [S.l.]: Springer, 2013. p. 37–52.

SMITH, M. R.; MARTINEZ, T.; GIRAUD-CARRIER, C. An instance level analysis of data complexity. Machine Learning, v. 95, n. 2, p. 225–256, 2014.

SOARES, R. V. Determinação de um índice de perigo de incêndio para a região centro-paranaense, Brasil. Dissertação (Dissertação de Mestrado) — Centro Tropical de Ensino e Investigação, Instituto Interamericano de Ciências Agrícolas, Costa Rica, 1972.

SOARES, R. V.; BATISTA, A. C. Incêndios florestais: controle, efeitos e uso do fogo. [S.l.: s.n.], 2007. ISBN 9788590435327.

SORIANO, B. M. A.; DANIEL, O.; SANTOS, S. A. Eficiência de índices de risco de incêndios para o pantanal sul-mato-grossense. Ciência Florestal, v. 25, n. 4, p. 809–816, 2015. ISSN 0103-9954.

TAN, P.-N.; STEINBACH, M.; KUMAR, V. Introduction to Data Mining, (First Edition). USA: Addison-Wesley Longman Publishing Co., Inc., 2005. ISBN 0321321367.

TELICYN, G. P. Logarithmic index of fire weather danger for forests. Lesnoe Khozyaistvo, v. 11, n. 1, p. 1–58, 1970.

TETTO, A. F. et al. Comportamento e ajuste da fórmula de monte alegre na floresta nacional de Irati, estado do Paraná. Scientia Forestalis, v. 38, n. 87, p. 409–417, 2010.

TORRES, F. T. P.; LIMA, G. S. Conservation of nature forest fire hazard in the Serra do Brigadeiro State Park (MG). Floresta e Ambiente (FLORAM), v. 26, n. 2, 2019.

TORRES, F. T. P. et al. Analysis of efficiency of fire danger indices in forest fire prediction. Revista Árvore, v. 41, n. 2, 2017. TORRES, F. T. P.; RIBEIRO, G. A. índices de risco de incêndios florestais em Juiz de Fora/MG. Floresta e Ambiente (FLORAM), v. 15, n. 2, p. 24–34, 2008.

VIGANÓ, H. H. da G. et al. Redes neurais artificiais na previsão de queimadas e incêndios no Pantanal. Revista Brasileira de Geografia Física, 2017. ISSN 1984-2295.

WAGNER, C. V. Development and structure of the canadian forest fire weather index system. Canadian Forest Service Publications, v. 35, 1987.

WILLMOTT, C. J.; MATSUURA, K. Advantages of the mean absolute error (mae) over the root mean square error (rmse) in assessing average model performance. Climate Research, v. 30, n. 1, 2005.

YAMASHITA, R. et al. Convolutional neural networks: an overview and application in radiology. Insights into Imaging, v. 9, n. 4, p. 611–629, 2018.

YANG, S.; LUPASCU, M.; MEEL, K. S. Predicting forest fire using remote sensing data and machine learning. Computer Vision and Pattern Recognition (CoRR), 2021.

ZHANG, J.; MANI, I. KNN Approach to Unbalanced Data Distributions: A Case Study Involving Information Extraction. In: Proceedings of the ICML'2003 Workshop on Learning from Imbalanced Datasets. [S.l.: s.n.], 2003.