REVISTA
DESAFIOS
ISSN: 2359-3652
V.13,n.3, maio/2026–DOI:10.20873/vol13n3pibic20258

ANÁLISE CLIMÁTICA NO ESTADO DO TOCANTINS UTILIZANDO APRENDIZADO DE MÁQUINA

CLIMATE ANALYSIS IN THE STATE OF TOCANTINS USING MACHINE LEARNING

ANÁLISIS CLIMÁTICO EN EL ESTADO DE TOCANTINS UTILIZANDO APRENDIZAJE AUTOMÁTICO

José Lucas Carvalho Silva

lucas.jose1@uft.edu.br

Glenda Michele Botelho

glendabotelho@uft.edu.br

ABSTRACT:
This study applied data mining techniques to climate records from Tocantins state, Brazil, to develop predictive models for precipitation, average temperature, and relative humidity. The CRISP-DM methodology and Random Forest algorithms were employed, based on data from five INMET meteorological stations distributed across the state, covering the period from 2010 to 2025. Data preprocessing involved cleaning, temporal aggregation, outlier treatment, feature engineering with temporal lags, moving averages, and cyclical encoding of seasonal variables. The dataset was split chronologically into training (2010–2021), validation (2022–2023), and test (2024–2025) sets. The models demonstrated robust performance, with coefficients of determination exceeding 0.94 for all variables in the test set. The humidity model achieved the best performance ($R^2 = 0.966$), followed by precipitation ($R^2 = 0.960$) and average temperature ($R^2 = 0.945$). Predictions for 2026 maintained fidelity to historical seasonal patterns, with seasonal correlations above 0.997 and approval in all six scientific validation criteria. Temperature predictions indicated regional climate stability, with an annual average of 26.2°C (0.1°C difference from historical climatology), with no significant warming or cooling trends. These results provide a scientifically validated predictive tool for agricultural planning and water resource management in Tocantins, supporting strategic decision-making in the state's agricultural sector.

KEYWORDS: Artificial Intelligence; Climate Prediction; Machine Learning; Random Forest; Tocantins.

RESUMO:
Este estudo aplicou técnicas de mineração de dados aos registros climáticos do estado do Tocantins com o objetivo de desenvolver modelos preditivos para precipitação, temperatura média e umidade relativa. Foram empregados a metodologia CRISP-DM e o algoritmo *Random Forest*, a partir de dados de cinco estações meteorológicas do INMET distribuídas pelo estado, cobrindo o período de 2010 a 2025. O pré-processamento envolveu limpeza, agregação temporal, tratamento de *outliers* e engenharia de atributos com defasagens temporais, médias móveis e codificação cíclica de variáveis sazonais. Os dados foram divididos cronologicamente em treino (2010–2021), validação (2022–2023) e teste (2024–2025). Os modelos demonstraram desempenho robusto, com coeficientes de determinação superiores a 0,94 em todas as variáveis no conjunto de teste. O modelo de umidade apresentou o melhor resultado ($R^2 = 0{,}966$), seguido pela precipitação ($R^2 = 0{,}960$) e temperatura média ($R^2 = 0{,}945$). As predições para 2026 mantiveram fidelidade aos padrões sazonais históricos, com correlações sazonais superiores a 0,997 e aprovação em todos os seis critérios de validação científica. As predições de temperatura indicaram estabilidade climática regional, com média anual de 26,2°C (diferença de 0,1°C da climatologia histórica), sem tendências significativas de aquecimento ou resfriamento. Os resultados fornecem uma ferramenta preditiva validada cientificamente para o planejamento agrícola e a gestão de recursos hídricos no Tocantins, contribuindo para a tomada de decisões estratégicas no setor agropecuário estadual.

PALAVRAS-CHAVE: Inteligência Artificial; Predição Climática; Aprendizado de Máquina; *Random Forest*; Tocantins.

RESUMEN:
Este estudio aplicó técnicas de minería de datos a los registros climáticos del estado de Tocantins, Brasil, con el objetivo de desarrollar modelos predictivos para precipitación, temperatura media y humedad relativa. Se emplearon la metodología CRISP-DM y algoritmos Random Forest, con base en datos de cinco estaciones meteorológicas del INMET distribuidas por el estado, abarcando el período de 2010 a 2025. El preprocesamiento incluyó limpieza, agregación temporal, tratamiento de valores atípicos e ingeniería de características con rezagos temporales, promedios móviles y codificación cíclica de variables estacionales. Los datos fueron divididos cronológicamente en entrenamiento (2010–2021), validación (2022–2023) y prueba (2024–2025). Los modelos demostraron un desempeño robusto, con coeficientes de determinación superiores a 0,94 para todas las variables en el conjunto de prueba. El modelo de humedad presentó el mejor rendimiento ($R^2 = 0,966$), seguido por precipitación ($R^2 = 0,960$) y temperatura media ($R^2 = 0,945$). Las predicciones para 2026 mantuvieron fidelidad a los patrones estacionales históricos, con correlaciones superiores a 0,997 y aprobación en los seis criterios de validación científica. Los resultados ofrecen una herramienta predictiva científicamente validada para la planificación agrícola y la gestión de recursos hídricos en Tocantins, contribuyendo a la toma de decisiones estratégicas en el sector agropecuario estatal.

PALABRAS CLAVE: Inteligencia Artificial; Predicción Climática; Aprendizaje Automático; Random Forest; Tocantins.

INTRODUÇÃO

A análise e compreensão das mudanças climáticas são questões cruciais para a sustentabilidade e o desenvolvimento de qualquer região. No contexto brasileiro, especialmente na vasta extensão do Cerrado, essas preocupações são ainda mais prementes, dada a sua riqueza em biodiversidade e sua importância para o equilíbrio ambiental. Segundo o *Intergovernmental Panel on Climate Change* (IPCC), ao longo do último século, o acúmulo de Gases do Efeito Estufa (GEEs) tem sido uma ameaça para a diversidade biológica no Cerrado, evidenciando que as mudanças climáticas produzem impactos mesmo em regiões com pouca interferência humana (IPCC, 2013).

O estado do Tocantins, localizado na região Norte do Brasil, não escapa a essas dinâmicas climáticas. A compreensão de sua dinâmica climática é essencial não apenas para a preservação do ambiente, mas também para subsidiar o planejamento e a organização das atividades econômicas, em especial as agrícolas. Por meio da avaliação de modelos climáticos, órgãos estaduais e instituições de pesquisa podem prover suporte para a tomada de decisões estratégicas. Nesse contexto, as projeções climáticas tornam-se ferramentas vitais para o planejamento e gestão hidroclimatológica.

Especialmente para o Brasil, cuja matriz energética é fortemente centrada na hidroeletricidade, a capacidade de estruturar modelos climáticos que prognostiquem possíveis impactos futuros é essencial para o planejamento energético de médio prazo. A preocupação com as mudanças climáticas também se estende aos eventos extremos, que estão intrinsecamente ligados aos fenômenos El Niño e La Niña. Esses padrões climáticos, referentes ao aquecimento e resfriamento anormal das águas do Oceano Pacífico equatorial, respectivamente, podem desencadear extremos climáticos significativos. Tais eventos podem ter ramificações sociais,

econômicas e ambientais substanciais, afetando desde o conforto humano até o turismo e o consumo de energia (DIAS, 2014; QIAN; LIN, 2005; SANTOS et al., 2009).

O estado do Tocantins, com 50,25% de sua superfície adequada para a agricultura, destaca-se como um novo polo agrícola do Brasil (SEAGRO, 2017). No entanto, as atividades agrícolas dependem intimamente das variáveis climáticas, cuja flutuação pode impactar significativamente os cultivos. Atualmente, o avanço da era da informação impulsiona a captação e armazenamento de diversas informações climáticas em vastas bases de dados, cujo crescimento contínuo resulta em uma enorme quantidade de registros e atributos.

Dentro desse contexto, torna-se crucial a identificação de padrões e relações ocultas nesses dados, tarefa praticamente impossível de ser realizada manualmente devido ao volume. A mineração de dados, aplicada em diversas áreas incluindo a climatologia, destaca-se como um procedimento automatizado para descobrir padrões úteis previamente ocultos em grandes bases de dados.

Este trabalho propõe-se a aplicar técnicas de mineração de dados aos registros climáticos do estado do Tocantins, visando desenvolver modelos preditivos para as principais variáveis climáticas da região. Essa abordagem permitirá a predição de precipitação, temperatura e umidade em diferentes escalas temporais, com o objetivo de auxiliar na tomada de decisões estratégicas e na implementação de medidas preventivas para o setor agropecuário do estado.

Diante desse panorama, é crucial o desenvolvimento de sistemas computacionais capazes de analisar os fatores e variáveis climáticas para oferecer resultados confiáveis sobre as mudanças no clima regional. Isso não apenas permite minimizar problemas ambientais e socioeconômicos decorrentes de eventos climáticos, mas também proporciona uma base sólida para o planejamento agrícola e econômico. Esses processos permitem a transformação de vastas quantidades de dados em conhecimento acionável, crucial para o processo de tomada de decisões (KOH; TAN et al., 2011). Neste contexto, este estudo emprega a metodologia CRISP-DM e algoritmos de aprendizado de máquina para realizar predições climáticas em diferentes escalas temporais, incluindo anual, mensal e diária.

METODOLOGIA

Para a execução do estudo, utilizou-se a metodologia CRISP-DM (*Cross Industry Standard Process for Data Mining*), amplamente aplicada em projetos de mineração de dados. De acordo com Wirth e Hipp (2000), o processo CRISP-DM é composto por seis etapas fundamentais: compreensão do negócio, compreensão dos dados, preparação dos dados, modelagem, avaliação e implantação. A metodologia CRISP-DM permite um fluxo interativo de análise, garantindo refinamento contínuo dos dados e modelos empregados.

1. ÁREA DE ESTUDO E DADOS

Neste trabalho, a área de estudo compreende o estado do Tocantins, localizado entre as longitudes 45ºW e 51ºW e latitudes 5ºS e 14ºS. Os dados climáticos utilizados foram obtidos a partir das séries históricas disponibilizadas pelo Instituto Nacional de Meteorologia (INMET), abrangendo um período de 21 anos, de 2004 a 2025. Para garantir uma representação espacial adequada das diferentes condições climáticas regionais, foram selecionadas cinco estações meteorológicas estrategicamente distribuídas no estado: Araguaína, Palmas, Pedro Afonso, Dianópolis e Gurupi.

As variáveis consideradas na pesquisa incluem: data, precipitação total (mm), temperatura máxima (°C), temperatura média (°C), temperatura mínima (°C), umidade relativa média (%), velocidade do vento média (m/s) e pressão atmosférica (mB).

2. METODOLOGIA CRISP-DM APLICADA AO ESTUDO

2.1 COMPREENSÃO DO NEGÓCIO

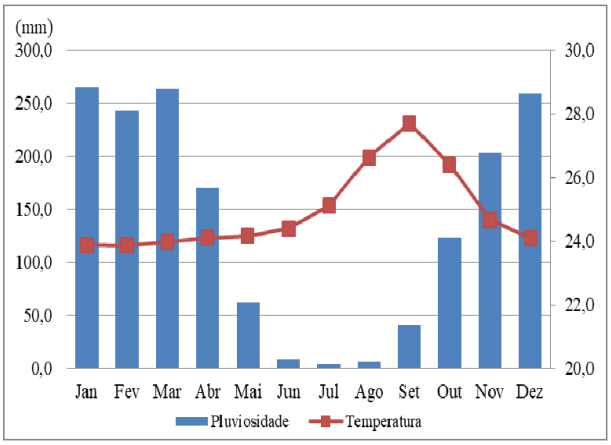
A pesquisa utilizou séries de dados contínuas e homogêneas, selecionando exclusivamente estações meteorológicas com disponibilidade e qualidade satisfatórias em seus registros. O período considerado abrange de 2004 a 2025, contemplando apenas estações telemétricas. Os dados meteorológicos foram coletados em escala diária para garantir que a agregação mensal refletisse corretamente os valores reais, evitando distorções decorrentes de dias faltantes nas análises. Nem todas as estações possuíam dados desde 2004, pois o início das séries está associado à data de instalação — algumas iniciaram em 2006, 2007 ou 2008. As séries históricas foram extraídas do sistema do INMET em formato CSV.

2.2 COMPREENSÃO DOS DADOS

O estado do Tocantins apresenta características climáticas típicas do clima tropical, com duas estações bem definidas: uma estação chuvosa (outubro a março) concentrando aproximadamente 80% da precipitação anual, e uma estação seca (abril a setembro). Janeiro é o mês mais chuvoso e julho o mais seco. A pluviosidade anual varia entre 1.200 a 2.000 mm, com maiores precipitações nas regiões norte e centro (ROLDÃO; FERREIRA, 2019).

O regime térmico caracteriza-se por temperaturas médias anuais entre 24°C e 28°C, com amplitude térmica baixa. O trimestre mais quente (agosto-outubro) antecede a estação chuvosa, sendo setembro o mês mais quente e janeiro o mais frio. Esta distribuição reflete a influência da sazonalidade pluviométrica sobre o regime térmico regional, evidenciando o padrão climático tropical continental brasileiro (ROLDÃO; FERREIRA, 2019). O Gráfico 1 ilustra o climograma estadual (1985-2016), evidenciando o padrão bimodal característico.

Gráfico 1 – Climograma do Estado do Tocantins (1985 a 2016).



Fonte: ANA (2018); NCEP/NCAR Reanalysis (2018). Fonte: ROLDÃO; FERREIRA (2019).

2.3 PREPARAÇÃO DOS DADOS

A preparação dos dados constitui uma etapa fundamental de qualquer investigação científica baseada em dados, pois assegura a qualidade, consistência e integridade das informações utilizadas no estudo. Para este trabalho, os arquivos CSV correspondentes a cada cidade foram consolidados em um único conjunto de dados, totalizando 34.043 registros. As principais etapas de pré-processamento incluíram: limpeza dos dados, conversão de tipos e padronização de formatos, tratamento de *outliers*, agregação temporal e codificação de variáveis categóricas.

2.3.1 LIMPEZA DOS DADOS, CONVERSÃO DE TIPOS E FORMATAÇÃO

A etapa de limpeza dos dados é essencial para assegurar a consistência, integridade e confiabilidade das informações antes da análise estatística ou aplicação de modelos de aprendizado de máquina. Os tipos de dados foram padronizados: a variável temporal 'Data' foi convertida para o formato datetime, as variáveis meteorológicas foram transformadas em tipo numérico, e a variável categórica 'Cidade' foi convertida para categoria. Registros duplicados foram removidos com base na combinação única de 'Data' e 'Cidade'.

Para o tratamento de dados faltantes, foi adotada uma abordagem hierárquica. Valores ausentes de precipitação foram imputados como zero, seguindo a prática hidrometeorológica em que a ausência de registro geralmente indica ausência de chuva (BAUER et al., 2018; KIDD et al., 2008). Para as demais variáveis meteorológicas, adotou-se primeiro a interpolação linear para gaps curtos (até três dias consecutivos), técnica recomendada para séries climáticas com autocorrelação devido à sua eficácia em reduzir erros em lacunas curtas (ANDERSON, 2018). Em seguida, utilizou-se a média sazonal por mês e cidade, preservando o padrão climático cíclico conforme práticas recomendadas (AL-ANSARI et al., 2023). Como último recurso, aplicou-se a média geral histórica da cidade. Por fim, foram realizadas validações de consistência lógica,

verificando a hierarquia térmica, limites físicos de umidade relativa (0–100%) e ausência de precipitação negativa.

2.3.2 TRATAMENTO DE OUTLIERS

A detecção e remoção de *outliers* foi conduzida considerando tanto critérios estatísticos quanto a plausibilidade física dos dados meteorológicos. O método do intervalo interquartil (IQR), proposto por Tukey (1977), foi utilizado para identificar valores atípicos em variáveis como temperatura e umidade relativa, pois , sua aplicação direta pode levar à exclusão de eventos extremos genuínos, que são cientificamente relevantes. Pensando nisso, para precipitação, optou-se por utilizar um limite físico de plausibilidade, em consonância com as recomendações da Organização Meteorológica Mundial (WMO, 2017), estabelecendo-se o limite de 150 mm/dia como valor máximo plausível. A variável pressão foi removida da base, pois apresentava valores que resultam em exclusão excessiva de registros quando aplicado o método IQR, e não era essencial para os objetivos do estudo. Essa abordagem, combinando limites estatísticos e físicos, está em consonância com práticas de controle de qualidade de dados climatológicos em regiões tropicais, onde a variabilidade natural é elevada. Estudos apontam para a importância de preservar eventos extremos como parte fundamental da variabilidade climática (WMO, 2017).

2.3.3 AGREGAÇÃO TEMPORAL

A etapa de agregação temporal foi realizada com o objetivo de transformar as séries diárias em registros mensais, assegurando maior robustez estatística e permitindo a análise de padrões sazonais de forma consistente com a literatura climatológica. Para a variável precipitação, adotou-se a soma mensal, uma vez que esta representa o acúmulo natural do fenômeno ao longo do tempo; já para a temperatura média e a umidade relativa, optou-se pelo cálculo da média mensal, refletindo adequadamente as condições médias de cada período.

O recorte temporal utilizado compreendeu os anos de 2010 a 2025, selecionados por apresentarem maior consistência e completude dos registros, garantindo representatividade climatológica sem vieses decorrentes de falhas em séries históricas anteriores. Durante o processo, identificaram-se valores anômalos de precipitação igual a zero em meses tipicamente chuvosos, como dezembro e janeiro; tais registros foram tratados mediante substituição pela mediana dos valores mensais válidos da mesma localidade, preservando a coerência com o regime pluviométrico regional (ROLDÃO; FERREIRA, 2019).

2.3.4 CODIFICAÇÃO DE VARIÁVEIS CATEGÓRICAS

A transformação da variável Cidade em formato numérico foi necessária para viabilizar o uso de algoritmos de aprendizado de máquina, que trabalham apenas com dados quantitativos. Para isso, empregou-se o método Label Encoding, disponibilizado pela biblioteca Scikit-learn (PEDREGOSA et al., 2011). Esse procedimento associa um número inteiro a cada cidade, de forma única e consistente, sem introduzir qualquer ordem artificial entre as categorias. A escolha

por esse método deve-se à sua simplicidade, boa adequação a variáveis nominais e ampla aceitação em tarefas de pré-processamento de dados.

2.3.5 SELEÇÃO E CRIAÇÃO DE ATRIBUTOS

A engenharia de atributos é uma das etapas mais determinantes para o desempenho em tarefas de predição, pois incorpora conhecimento de domínio no processo de modelagem (DOMINGOS, 2012; KUHN; JOHNSON, 2019). A etapa de seleção e criação de atributos foi essencial para ampliar a capacidade explicativa do conjunto de dados e fornecer ao modelo variáveis mais representativas dos fenômenos climáticos. Além das variáveis originais (precipitação, temperatura média e umidade), foram desenvolvidos atributos de defasagem temporal (lags), médias móveis de curto e médio prazo, indicadores sazonais e codificações cíclicas do calendário (como seno e cosseno do mês), preservando a natureza periódica dos dados climáticos.

Também foram criadas variáveis derivadas, como diferenças entre meses consecutivos e desvios em relação às tendências de longo prazo, além de uma interação cidade-mês que capta padrões sazonais específicos por localidade. Essa estratégia permitiu que os algoritmos de aprendizado de máquina explorassem tanto dependências temporais de curto prazo quanto padrões sazonais e espaciais mais complexos, aumentando a robustez do modelo.

2.3.6 DIVISÃO DOS DADOS

A divisão temporal dos dados desempenhou papel estratégico na estruturação do estudo, assegurando que o processo de modelagem respeitasse tanto a natureza sequencial das séries temporais climáticas quanto a necessidade de generalização dos algoritmos. Para isso, os dados foram organizados em três blocos cronológicos: treino (2010–2021), validação (2022–2023) e teste (2024–2025). Essa segmentação buscou equilibrar a robustez estatística com a representatividade climática, contemplando diferentes condições sazonais e interanuais.

Ao mesmo tempo, garantiu-se a simulação fiel de cenários preditivos reais, nos quais apenas informações passadas podem ser utilizadas para prever valores futuros. Essa escolha metodológica segue recomendações consolidadas em estudos sobre avaliação de modelos de séries temporais, que destacam a importância da divisão cronológica como forma de reduzir vieses e prevenir vazamento de dados (BERGMEIR; HYNDMAN; KOO, 2018).

2.3.7 TREINAMENTO DO MODELO

A etapa de treinamento foi conduzida de forma independente para cada variável climática alvo — precipitação, temperatura média e umidade — utilizando três modelos *Random Forest* de regressão. A escolha deste algoritmo deve-se à sua robustez em problemas de previsão com múltiplas variáveis e séries temporais, conforme discutido por Breiman (2001). Os modelos foram configurados com parâmetros otimizados para o conjunto de dados disponível: 100

estimadores, profundidade máxima de 10 níveis, mínimo de 5 amostras por divisão e 2 por folha, seleção de características pelo critério sqrt e validação interna via out-of-bag.

O treinamento considerou exclusivamente o período de 2010 a 2021, incorporando todas as características temporais, sazonais e derivadas previamente construídas. Para lidar com dados ausentes, aplicou-se imputação pela mediana, estratégia considerada adequada para variáveis ambientais sujeitas a *outliers* (LÓPEZ-ROMERO et al., 2018).

A avaliação dos modelos foi conduzida por meio de um conjunto de métricas complementares: RMSE (Root Mean Squared Error), MAE (Mean Absolute Error), R^2 (coeficiente de determinação) e MAPE (Mean Absolute Percentage Error), aplicadas nas etapas de treino, validação e teste, com o objetivo de obter uma análise abrangente do desempenho. O RMSE mede a raiz do erro quadrático médio, penalizando grandes desvios entre valores previstos e observados, sendo sensível à presença de outliers. O MAE, por sua vez, representa o erro absoluto médio, fornecendo uma medida direta e interpretável da magnitude média dos erros. Já o R^2 indica a proporção da variância dos dados explicada pelo modelo, permitindo avaliar o quanto as previsões se aproximam do comportamento real, com valores mais próximos de 1 indicando melhor ajuste. Por fim, o MAPE expressa o erro percentual médio, facilitando a interpretação relativa do desempenho do modelo em termos percentuais..

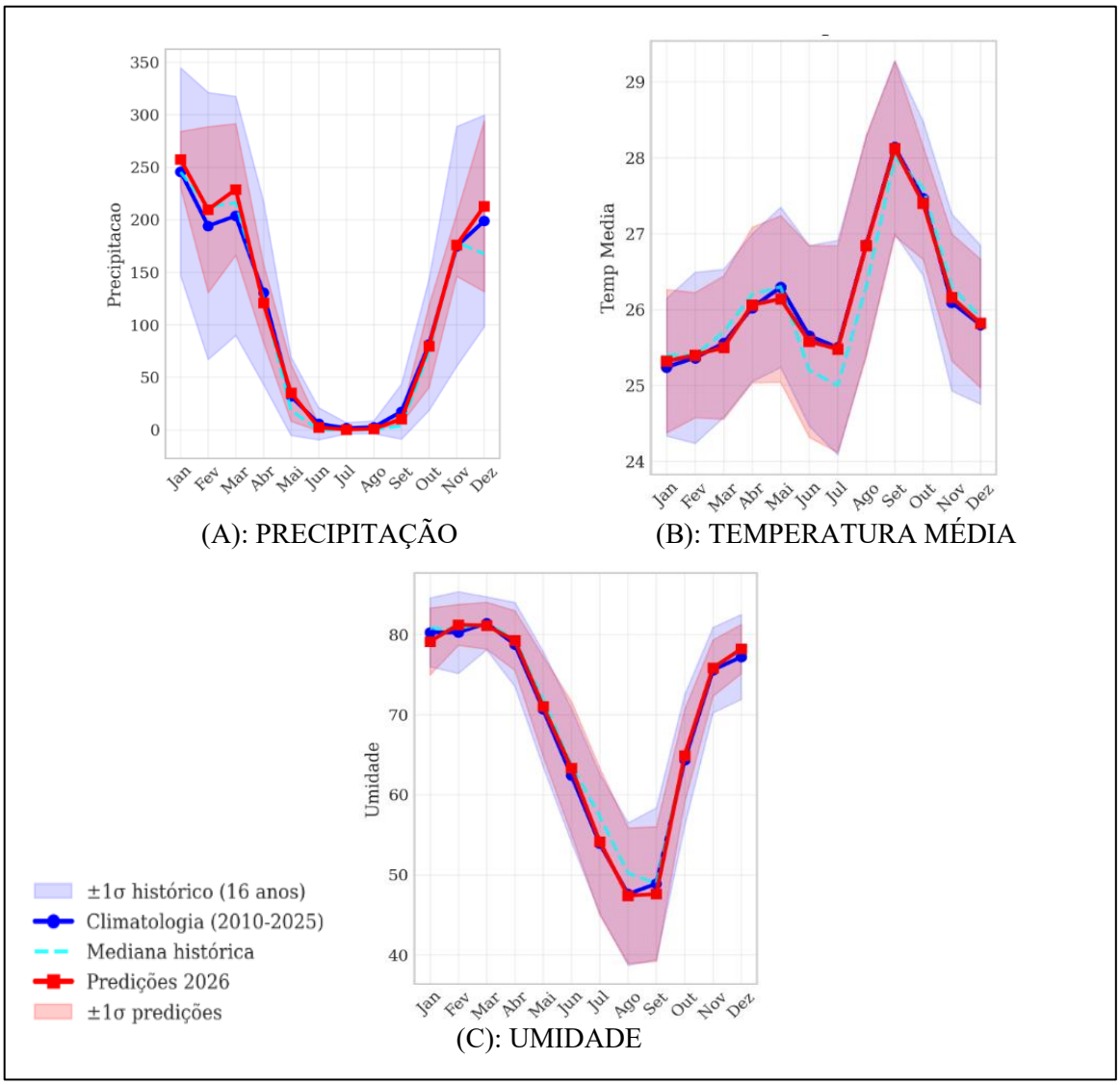
RESULTADOS E DISCUSSÃO

A análise comparativa entre as predições climáticas para 2026 e a climatologia histórica de 16 anos (2010-2025) demonstrou consistência com os padrões sazonais estabelecidos para o estado do Tocantins. As predições mantiveram a característica bimodal do clima tropical regional, com estação chuvosa concentrada entre dezembro e março e estação seca de junho a agosto (ROLDÃO; FERREIRA, 2019).

Para precipitação, o modelo preservou janeiro como mês mais chuvoso (predito: janeiro, histórico: janeiro) e julho como mais seco (predito: julho, histórico: julho), demonstrando correlação sazonal de 0,997 ($p<0,001$). As predições de temperatura média para 2026 demonstraram alta estabilidade climática, com média anual de 26,2°C, praticamente idêntica à climatologia histórica de 16 anos (26,1°C), representando diferença de apenas 0,1°C. Não foram identificadas tendências significativas de aquecimento ou resfriamento em relação ao período base (2010-2025), indicando estabilidade térmica regional apropriada para predições de curto prazo.

As temperaturas seguiram o padrão sazonal de máximas em setembro e mínimas em janeiro, enquanto a umidade reproduziu fielmente a variação sazonal inversa à temperatura. Esta consistência valida a capacidade dos modelos em capturar e projetar adequadamente os padrões climáticos regionais estabelecidos. A análise comparativa entre as predições climáticas para 2026 e a climatologia histórica é apresentada no Gráfico 2, que demonstra a consistência dos modelos com os padrões sazonais estabelecidos.

Gráfico 2 – Comparação entre predições 2026 e climatologia histórica para precipitação, temperatura média e umidade relativa.



(A): PRECIPITAÇÃO

(B): TEMPERATURA MÉDIA

(C): UMIDADE

Fonte: Elaboração própria.

3.1 PERFORMANCE DOS MODELOS RANDOM FOREST

Os modelos Random Forest desenvolvidos para cada variável climática (precipitação, temperatura média e umidade) demonstraram performance robusta na predição das variáveis climáticas do Tocantins, com coeficientes de determinação superiores a 0,94 em todos os casos no conjunto de teste. O modelo de umidade apresentou o melhor desempenho (R^2 = 0,966), seguido pela precipitação (R^2 = 0,960) e temperatura média (R^2 = 0,945), resultando em R^2 médio de 0,957 para o conjunto de modelos.

A análise das métricas de erro revelou comportamentos distintos entre as variáveis. O modelo de temperatura média apresentou os menores erros absolutos (RMSE = 0,32°C, MAE = 0,25°C), refletindo a maior estabilidade temporal dessa variável. O modelo de precipitação,

embora mantendo alta correlação, apresentou maiores variações absolutas (RMSE = 22,3 mm), comportamento esperado devido à natureza mais variável e intermitente da precipitação. Estudos demonstram consistentemente que a temperatura é prevista com maior precisão que a precipitação, com modelos Random Forest tipicamente alcançando RMSE entre 40–51 mm para precipitação mensal. O modelo de umidade demonstrou desempenho intermediário (RMSE = 2,6%), confirmando a capacidade dos algoritmos em capturar adequadamente a dinâmica das três variáveis climáticas principais.

Tabela 1 – Performance dos modelos Random Forest para precipitação, temperatura média e umidade relativa nos conjuntos de treino, validação e teste.

Variável	R^2 (Treino)	R^2 (Validação)	R^2 (Teste)	RMSE (Teste)	MAE (Teste)	MAPE (Teste)
Precipitação	0.983	0.919	0.960	22.29 mm	17.22 mm	127.9%
Temperatura	0.985	0.947	0.945	0.32°C	0.25°C	1.0%
Umidade	0.992	0.974	0.966	2.62%	2.05%	3.1%

Fonte: Elaboração própria.

3.2 PREDIÇÕES CLIMÁTICAS PARA 2026

As predições climáticas para 2026 foram baseadas em série histórica de 16 anos (2010-2025) e submetidas a seis critérios de validação: range físico válido, diferença da climatologia, preservação da variabilidade, correlação sazonal, controle de extremos e hierarquia sazonal. Todos os modelos foram aprovados em todos os critérios (100%), confirmando a robustez científica das predições.

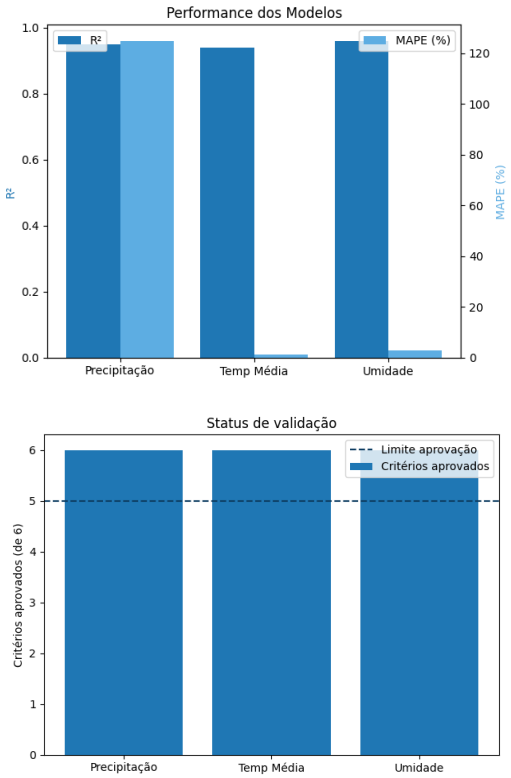
Para precipitação, as predições indicam padrão anual médio de 111,1 mm/mês (diferença de 3,3% da climatologia histórica), preservando a sazonalidade regional com mínimos em julho (1,1 mm) e máximos no trimestre chuvoso. A variabilidade predita (105,2 mm) mostrou-se consistente com a natural observada (119,8 mm). As predições de temperatura média demonstraram alta precisão, com média anual de 26,2°C (climatologia: 26,1°C) e range de 23,8–29,7°C, mantendo o padrão sazonal com máximas em setembro e mínimas em janeiro.

A umidade relativa predita (68,6%) reproduziu fielmente o comportamento histórico, preservando a variação sazonal inversa à temperatura. A validação científica atestou correlações sazonais superiores a 0,997 para todas as variáveis, ausência de valores extremos anômalos (frequência < 10%) e manutenção da hierarquia sazonal histórica, confirmando que as predições são climatologicamente consistentes. As Figuras 3 e 4 apresentam, respectivamente, a performance dos modelos Random Forest e o status de validação científica das predições, confirmando a robustez dos resultados obtidos.

O MAPE da precipitação (127,9%) apresenta valor elevado, porém isso não invalida o modelo. Essa métrica deve ser interpretada com cautela, uma vez que a precipitação frequentemente assume valores próximos de zero, condição na qual o MAPE tende a inflar

significativamente os erros percentuais devido à divisão pelo valor observado. Trata-se de uma limitação discutida na literatura de avaliação de modelos preditivos, especialmente em séries com distribuição assimétrica e presença de zeros, nas quais o MAPE pode produzir valores distorcidos e pouco representativos do desempenho do modelo (HYNDMAN; KOEHLER, 2006; KIM; KIM, 2016).

Gráficos 3 e 4 – Performance dos modelos Random Forest e validação científica das predições.



Fonte: Elaboração própria.

CONSIDERAÇÕES FINAIS

Este estudo aplicou com sucesso técnicas de mineração de dados aos registros climáticos do estado do Tocantins, desenvolvendo modelos preditivos robustos para precipitação, temperatura média e umidade relativa através da metodologia CRISP-DM e algoritmos Random Forest. Os modelos desenvolvidos demonstraram performance consistente, com coeficientes de determinação superiores a 0.94 para todas as variáveis no conjunto de teste. O modelo de umidade apresentou o melhor desempenho ($R^2 = 0.966$), seguido pela precipitação ($R^2 = 0.960$) e temperatura média ($R^2 = 0.945$). As predições climáticas para 2026 mantiveram fidelidade aos padrões sazonais históricos, com correlações superiores a 0.997 e aprovação em todos os critérios de validação científica. A principal contribuição deste trabalho consiste no desenvolvimento de uma ferramenta preditiva validada cientificamente para o planejamento agrícola e gestão de recursos hídricos no Tocantins. Os resultados fornecem subsídios técnicos para tomada de decisões estratégicas no setor agropecuário, considerando a importância econômica da agricultura para o estado.

Agradecimentos

O presente trabalho foi realizado com o apoio do Conselho Nacional de Desenvolvimento Científico e Tecnológico — CNPq — Brasil.

REFERÊNCIAS

AL-ANSARI, N. et al. Univariate and multivariate imputation methods evaluation for reconstructing climate time series data: A case study of Mosul station-Iraq. **Frontiers in Earth Science**, v. 11, 2023.

ANA – Agência Nacional de Águas. **Hidroweb: sistema de informações hidrológicas**. Brasília: ANA, 2018. Disponível em: http://hidroweb.ana.gov.br.

ANDERSON, T. M. On the effectiveness of interpolation methods for climate data gaps. **International Journal of Climatology**, v. 38, 2018. p. 1–18.

BAUER, P.; KIDD, C. et al. Satellite precipitation measurement. **Bulletin of the American Meteorological Society**, v. 99, n. 9, 2018. p. 1859–1886.

BERGMEIR, C.; HYNDMAN, R. J.; KOO, B. On the validity of cross-validation for evaluating time series prediction. **Journal of Forecasting**, v. 37, n. 8, 2018. p. 1–15.

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

DIAS, M. A. F. S. **Meteorologia: noções básicas**. São Paulo: Oficina de Textos, 2014.

DOMINGOS, P. A few useful things to know about machine learning. **Communications of the ACM**, v. 55, n. 10, 2012. p. 78–87.

IPCC. *Climate Change 2013: The Physical Science Basis*. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge: **Cambridge University Press**, 2013.

HYNDMAN, Rob J.; KOEHLER, Anne B. Another look at measures of forecast accuracy**. International Journal of Forecasting**, v. 22, n. 4, p. 679–688, 2006.

KIDD, C. et al. The Global Precipitation Measurement (GPM) mission. **Advances in Space Research**, v. 42, 2008. p. 1761–1774.

KIM, Sungil; KIM, Heung-Seok. A new metric of absolute percentage error for intermittent demand forecasts. **International Journal of Forecasting**, v. 32, n. 3, p. 669–679, 2016.

KOH, J.; TAN, C. et al. Data mining applications in sustainability. **Procedia Computer Science**, v. 4, 2011. p. 703–709.

KUHN, M.; JOHNSON, K. **Feature Engineering and Selection: A Practical Approach for Predictive Models**. Boca Raton: CRC Press, 2019.

LITTLE, R. J. A.; RUBIN, D. B. **Statistical analysis with missing data**. 2. ed. New York: Wiley, 2002.

LÓPEZ-ROMERO, E.; MORENO, C.; MARTÍN, P. Comparison of imputation methods for environmental time series. **Ecological Indicators**, v. 95, 2018. p. 1–10.

NCEP/NCAR. **Reanalysis Project Data**. National Centers for Environmental Prediction/National Center for Atmospheric Research, 2018. Disponível em: https://psl.noaa.gov/data/reanalysis.

PEDREGOSA, F. et al. Scikit-learn: Machine learning in Python. **Journal of Machine Learning Research**, v. 12, 2011. p. 2825–2830.

QIAN, W.; LIN, Z. Regional trends in recent precipitation indices in China. **Meteorology and Atmospheric Physics**, v. 90, 2005. p. 193–207.

ROLDÃO, G. R.; FERREIRA, N. J. Climatologia do Estado do Tocantins: variabilidade termo-pluviométrica no período de 1985–2016. **Revista Geonorte**, v. 10, n. 36, 2019. p. 155–178.

SANTOS, C. A. C.; BRAGA, C. C.; BRAGA, A. C. Climatologia do estado do Tocantins: variabilidade e tendências. **Revista Brasileira de Meteorologia**, v. 24, n. 1, 2009. p. 1–12.

SEAGRO – Secretaria da Agricultura, Pecuária e Abastecimento do Tocantins. **Relatório técnico anual 2017**. Palmas: SEAGRO, 2017.

WIRTH, R.; HIPP, J. CRISP-DM: Standard process for data mining. In: *Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining*. London, UK, 2000.

WMO – World Meteorological Organization. **Guidelines on quality control procedures for data from automatic weather stations**. Geneva: WMO, 2017.