Modelagem Preditiva de Variáveis Climáticas na Amazônia: Uma Análise Comparativa de Algoritmos de Machine Learning

Jhon Wallacy Virginia da Cruz,
José Jailton Henrique Ferreira Junior

[1]Faculdade de Computação (FACOMP) – Universidade Federal do Pará (UFPA)
Castanhal – PA – Brasil

jhon.cruz@castanhal.ufpa.br ,jjj@ufpa.br

Abstract. *This study aims to compare machine learning algorithms for forecasting meteorological characteristics such as precipitation and maximum temperature, using real-world data from the Amazon region. By analyzing performance metrics of the implemented models and considering the relationship between predictor variables and the target variable across different datasets, the study revealed critical insights for improving climate prediction models in the Amazon.*
Keywords: *Machine learning, Climate prediction, Amazon region, Precipitation, Maximum temperature.*

Resumo. *Este trabalho tem como objetivo, realizar uma comparação entre algoritmos de aprendizado de máquina para previsão de características meteorológicas como precipitação e temperatura máxima, tendo como conjunto de dados informações reais da região amazônica e analisar através de métricas de desempenho os modelos executados considerando a relação entre variáveis previsoras e variável prevista para diferentes conjuntos de dados. Dessa forma foi possível revelar informações importantes para aprimoramento de modelos de previsão climática na Amazônia.*
Palavras chave: *Aprendizado de máquina, Previsão climática, Região Amazônica, Precipitação, Temperatura máxima.*

1. Introdução

A previsão climática desempenha papel essencial em diversos setores da sociedade, como agricultura, geração de energia renovável e gestão de desastres naturais. Na Amazônia, região de extrema importância ecológica, a variabilidade climática se manifesta em extremos: de um lado, chuvas torrenciais causando alagamentos; de outro, ondas de calor com sensação térmica elevada, agravadas pela alta umidade. Eventos como a seca histórica de 2023 e projeções de redução de 25% na umidade relativa até 2040 alertam para uma crise multifacetada, onde a antecipação de riscos exige modelos capazes de prever tanto tempestades severas quanto o calor extremo [NASA 2024].

Desde meados do século XX, o monitoramento meteorológico tem sido fundamental para compreender as mudanças climáticas globais. A revolução começou com modelos numéricos baseados em equações diferenciais, como os desenvolvidos no

ENIAC em 1950, e evoluiu para sistemas que integram satélites, supercomputadores e técnicas estatísticas capazes de lidar com incertezas inerentes a fenômenos atmosféricos [Lezaun 2006].

Nos últimos anos, destaca-se o avanço da inteligência artificial e sua utilização para detecção de padrões a partir de uma vasta quantidade de dados gerados por máquina e por humanos, obtidos por meio de sensores e outros dispositivos inteligentes. Nesse contexto, os modelos de machine learning emergem como alternativas estratégicas, oferecendo vantagens ao monitoramento climático, pois são capazes de identificar relações não lineares entre variáveis como umidade, temperatura e velocidade do vento, relacionamentos essenciais para uma melhor representação de microclimas comuns na região amazônica [Artaxo et al. 2024].

Apesar da importância dos modelos de aprendizado de máquina para a previsão climática, a região Amazônica merece um destaque maior na literatura, visto que possui especificidades no clima.

A partir desse problema, este trabalho propõe uma análise comparativa de algoritmos de machine learning na previsão de variáveis climáticas como precipitação e temperatura em municípios da região amazônica. Para isso, foram utilizados dados meteorológicos do Instituto Nacional de Meteorologia [INMET 2025], submetidos a protocolos rigorosos de pré-processamento, incluindo tratamento de dados faltantes e seleção de features. Além de comparar o desempenho preditivo de modelos clássicos (ex.: regressão linear) e ensemble (ex.Random Forest, XGBoost e LightGBM), busca-se identificar relações entre variáveis climáticas e o desempenho dos modelos.

2. Referencial Teórico

2.1. Aprendizado de máquina

O aprendizado de máquina (ML) é um subcampo da inteligência artificial que se define por ser uma ciência que estuda algoritmos de programação capazes de adquirir capacidade de resolução em tarefas específicas por meio da experiência derivada de dados, sem intervenção humana explícita. Trata-se de um campo de estudo que dá aos computadores a habilidade de aprender sem serem explicitamente programados [Samuel 1959].

Essa abordagem parte da premissa de que sistemas computacionais podem identificar padrões complexos em dados históricos e generalizá-los para tomada de decisões futuras, um programa aprende com a experiência E em relação a uma tarefa T e uma métrica de desempenho P se sua eficácia em T, medida por P, melhora com E [Mitchell 1997].

Existem três paradigmas principais, quando tratamos de aprendizado de máquina [Faceli et al. 2021]:

1. Aprendizado supervisionado: Exige que os dados de treinamento fornecidos ao algoritmo inclua também as soluções desejadas, chamadas de rótulos.
2. Aprendizado não supervisionado: Utiliza dados não rotulados, sem resolução prévia, o próprio algoritmo será capaz de identificar estruturas e correlações entre os dados.
3. Aprendizado por reforço: Caracteriza-se pela existência de um agente capaz de observar o ambiente e executar ações que maximizem uma recompensa, muito eficiente em situações de ambiente dinâmico.

É definido que a eficiência de modelos de aprendizado de máquina depende de forma crítica da qualidade dos dados de treinamento e da seleção adequada de algoritmos, fatores que determinam diretamente a capacidade de generalização do sistema. Dados incompletos, ruidosos ou enviesados podem criar modelos que capturam padrões pouco significativos ou distorcidos, comprometendo aplicação em um cenário real. Além disso, a escolha inadequada de algoritmos mesmo que com dados de alta qualidade, podem resultar em subajuste (underfitting) ou superajuste (overfitting), nos quais o modelo falha em aprender relações significativas ou memoriza os dados de treinamento sem capacidade de generalizar para novas instâncias [Brownlee 2019].

2.2. Regressão

A regressão é uma técnica estatística que busca representar a relação entre uma ou mais variáveis independentes a uma variável dependente, permitindo prever valores com base em padrões implícitos na disposição dos dados [Soto 2013]. Uma variável independente é normalmente representada pela letra x, é um parâmetro que não sofre influência de outras variáveis em um determinado contexto. Já a variável dependente, que é normalmente representada pela letra y possui valor que está diretamente ligado à variável independente, de forma direta ou indireta conseguimos afirmar que x possui influência sobre y.

2.2.1. Regressão Linear

Quando aplicamos regressão ao âmbito de aprendizado de máquina supervisionado o algoritmo mais fundamental é a regressão linear, que é amplamente empregado para modelar relações lineares entre variáveis independentes e uma variável dependente de natureza contínua. Sua simplicidade matemática e interpretabilidade tornam esse método uma ferramenta essencial tanto para análises exploratórias quanto para previsões quantitativas em contextos práticos de natureza mais simples [Hastie et al. 2009].

Pode ser definida como uma função, de natureza linear que faz uma previsão calculando soma ponderada das informações de entrada mais uma constante que é chamada de termo de polarização (ou coeficiente linear) como mostrado na Equação (1).

$$\hat{y} = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_n x_n \tag{1}$$

- $\hat{y}$ é o valor a ser previsto;
- n é o número de entradas;
- x é o valor da *i-ésima* entrada;
- θ_j é o parâmetro do modelo j (incluindo o termo de polarização θ_0 e os pesos das entradas $\theta_1, \theta_2, \cdots, \theta_n$).

O objetivo do algoritmo é estimar os valores do parâmetro θ_j que é o coeficiente de polarização quando todas as variáveis independentes são zero e os valores dos coeficientes que representam os pesos de cada variável independente.

A equação pode ser expressa de maneira mais compacta utilizando uma forma vetorial, como mostrado na Equação (2).

$$\hat{y} = h_\theta(\mathbf{x}) = \theta^T \cdot \mathbf{x} \tag{2}$$

- θ é o vetor de parâmetro do modelo, que inclui o termo de polarização θ_0 e os pesos das entradas θ_1 a θ_n;
- θ^T é a transposição de θ
- x é o vetor de entradas da instância, que contém x_0 a x_n, com x_0 sempre igual a 1;
- h_0 é a função de hipótese, que utiliza os parâmetros do modelo θ.

2.2.2. Método dos Mínimos Quadrados

Para que seja realizado o treino em um modelo de regressão linear as estimativas de θ devem possuir valores de tal maneira que minimize a função de custo que representa o desempenho do treinamento. Ou seja, encontrar os melhores valores que permita que a diferença absoluta entre o valor esperado e o valor predito seja a menor possível entre todos os pontos.

Para que essa tarefa seja realizada temos como uma das opções o método dos mínimos quadrados, descrita na Equação (3), também conhecido como equação dos mínimos quadrados ordinários que é uma solução analítica e dá o resultado de forma direta. A equação combina álgebra linear e otimização para encontrar os coeficientes que melhor descrevem a relação linear entre as variáveis [Strang 2006].

$$\hat{\theta} = (X^\top \cdot X)^{-1} \cdot X^\top \cdot y \tag{3}$$

2.3. Métodos de Ensemble

Métodos de *ensemble* são técnicas mais avançadas entre os algoritmos de aprendizado de máquina supervisionado, pois combinam múltiplos modelos chamados de aprendizes fracos para produzir um modelo final mais robusto e preciso. Essa abordagem Segundo [Géron 2021] é baseada em um princípio ”conhecimento das multidões”, onde a diversidade de previsões de modelos individuais reduz erros e melhora a generalização.

2.3.1. Bagging e Boosting

Dentre as principais técnicas de *ensemble* temos *bagging* e *boosting* que de acordo com [Hastie et al. 2009] são técnicas baseadas no algoritmo de árvores de decisão apresentam bom tempo de processamento em fase de treinamento em comparação com outras técnicas de aprendizado de máquina.

A técnica de *bagging* atua utilizando o mesmo algoritmo para cada previsor de forma paralela, mas aplicando o mesmo treino a diferentes subconjuntos aleatórios do conjunto de treinamento, a previsão final consiste em coletar a média das saídas individuais dos previsores, reduzindo a variância do treinamento. Enquanto a técnica de *boosting* refere-se a ideia de treinar os previsores de maneira sequencial de modo que cada novo previsor atua na correção dos erros de seus predecessores, esse sequenciamento é realizado de forma iterativa convergindo ao final para um modelo forte, reduzindo o viés e trazendo precisão para dados complexos [Géron 2021].

3. Materiais e Métodos

Aqui serão abordados todos os aspectos metodológicos para a realização deste trabalho, descrevendo-se os procedimentos necessários e úteis para conseguir aplicar os algoritmos de aprendizado de máquina supervisionados para previsão de variáveis metereológicas, comparar o desempenho de diferentes algoritmos e avaliar correlações entre diferentes característas climáticas.

3.1. Coleta de dados

O conjunto de dados utilizados para construção dos modelos de machine learning neste trabalho foram retirados da base de dados do Instituto Nacional de Meteorologia [INMET 2025], órgão vinculado ao Ministério da Agricultura e Pecuária do Brasil, reconhecido por possuir a maior rede de estações automáticas da América do Sul. Todos os dados coletados por essa rede são disponibilizados gratuitamente em acesso público no portal online do INMET.

Os dados selecionados possuem frequência de registro diária em um período de janeiro de 2023 até fevereiro de 2025. Foram selecionados dados de 80 estações automáticas localizadas na região Norte do Brasil. A composição dos dados comuns a todas as estações podem ser observados na Tabela 1.

Tabela 1. Descrição das variáveis e suas unidades de medida

Variável	Unidade de Medida
Data	DD/MM/AAAA
Precipitação	mm
Pressão atmosférica	mbar
Ponto de orvalho	$^{\circ}\mathrm{C}$
Temperatura máxima	$^{\circ}\mathrm{C}$
Temperatura média	$^{\circ}\mathrm{C}$
Temperatura mínima	$^{\circ}\mathrm{C}$
Umidade relativa do ar média	$\%$
Umidade relativa do ar mínima	$\%$
Rajada de vento	$\mathrm{m/s}$
Velocidade do vento	$\mathrm{m/s}$

3.2. Tratamento de dados

Foi observado que a maioria das estações apresentava diversos registros diários com valores nulos, uma vez que estações automáticas podem apresentar mau funcionamento ou interrupção no processo de coleta e registro de informações meteorológicas. Para garantir a qualidade dos dados e evitar vieses nos modelos de machine learning, adotou-se um protocolo de seleção e tratamento.

Inicialmente, realizou-se a seleção das estações com menor incidência de dados ausentes. Estabeleceu-se um critério de retenção de até 15% de valores nulos por variável meteorológica, garantindo que todas as variáveis de interesse mantivessem consistência temporal. Estações que excederam esse limite em qualquer uma das variáveis foram excluídas, resultando em um subconjunto com 15 estações.

Para imputação dos dados remanescentes, aplicou-se uma abordagem temporal personalizada, considerando a natureza sequencial dos dados. Os métodos de preenchimento foram adaptados às características de cada variável, usando interpolação linear para variáveis contínuas como temperatura e preenchimento progressivo para variáveis como umidade, precipitação e velocidade do vento.

Como cada estação reflete o comportamento de um microclima distinto, a seleção das estações para treinamento dos modelos foi criteriosa no sentido de evitar criar muitos modelos. Dentre as 15 estações com dados completos, 5 foram selecionadas arbitrariamente, garantindo a representatividade dos microclimas, e utilizadas nos algoritmos de Regressão Linear, *Random Forest*, *XGBoost* e *LightGBM*.

3.3. Seleção de características para treinamento de modelos

Foi realizado uma análise exploratória com a agregação dos dados das cinco estações selecionadas com o objetico de observar as correlações lineares entre as variáveis para encontrar a melhor configuração de treinamento para a previsão de precipitação e temperatura máxima.

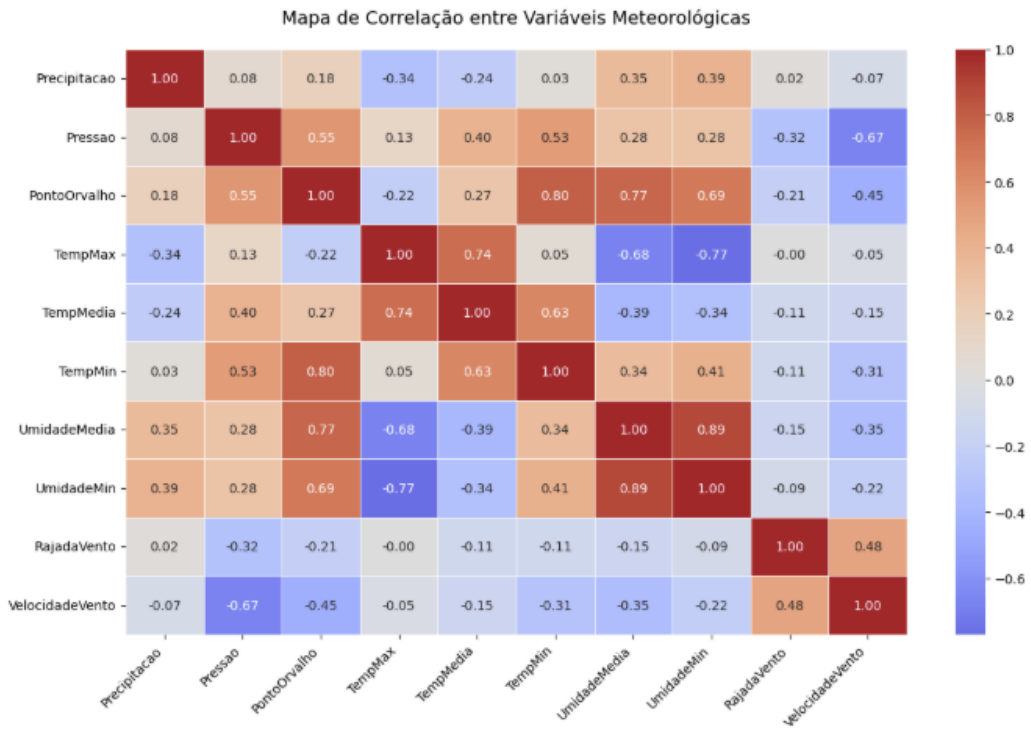

Figura 1. Mapa de correlação entre variáveis metereológicas
Fonte: De autoria própria.

Na Figura 1 temos o mapa que denota a correlação entre as váriaveis disponíveis para uso nos algoritmos, a partir desse mapa conseguimos avaliar a dependência linear entre as características e escolher o melhor conjunto de entradas para um modelo de previsão, além disso é possível observar variáveis que possuem correlação entre si muito fortes o que pode acabar trazendo redundância ao modelo de previsão, como é o caso das variáveis de temperatura máxima, média e mínima que possuem correlações fortes entre

si, dessa maneira ao selecionar as variáveis de entrada de um modelo é preciso eliminar a redundância [Berk 2008].

Considerando as informações de correlação foram selecionadas como características de entrada aos algortimos as variáveis de ponto de orvalho, temperatura máxima e umidade mínima para a previsão de precipitação. E para prever temperatura máxima foram selecionadas as características de pressão, umidade mínima e precipitação.

Além disso, foram criadas variáveis que representem informação de localização temporal, para que as séries sejam tratadas como séries temporais possibilitando aos algoritmos uma percepção de sazonalidade nos dados, estas variáveis representam dia no ano, dia na semana e mês.

As variáveis citadas constituem o conjunto de entradas base que será compartilhado por todos os modelos. Entretanto para fins experimentais e comparação de resultados também foram criadas variáveis temporais derivadas da variável a ser prevista, de modo a inputar informação histórica no treinamento dos modelos. Essa etapa foi realizada adicionando novas colunas de dados que representam de um até sete dias anteriores, além de uma coluna com o valor da média móvel dos últimos sete dias. Este novo conjunto adiciona ao treinamento dos modelos informações temporais extras.

4. Resultados

Nesta seção são apresentados os resultados obtidos com a execução dos diferentes modelos, denotando os valores de métricas como MAE (Erro Absoluto Médio), RMSE (Raiz do Erro Quadrático Médio) e R2 (Coeficiente de Determinação) que são métricas fundamentais para expressar a eficiência dos modelos de previsão.

Na execução do treinamento e validação dos modelos foram utilizados dois conjuntos diferentes como variáveis preditoras. Um dos conjuntos além das variáveis de base, foram utilizadas variáveis temporais derivadas da variável a ser prevista conforme apresentado na Seção 3.3 e o outro conjunto foram utilizadas apenas as variáveis de base.

No primeiro momento foram executados todos os algoritmos propostos para cada uma das estações utilizando o conjunto de variáveis base, com objetivo de prever a característica de precipitação, os resultados podem ser observados na Tabela 2.

Tabela 2. Desempenho dos modelos de previsão para precipitação sem uso de variáveis derivadas

Algoritmo	Estação	MAE	RMSE	R^2
Regressão Linear	Mateiros	4.61	8.88	0.25
	Belém	8.36	13.40	0.11
	Dom Eliseu	5.28	9.54	0.23
	C. do Araguaia	6.45	10.96	0.16
	Itaituba	7.29	11.90	0.21
Random Forest	Mateiros	4.23	8.87	0.25
	Belém	8.24	13.35	0.12
	Dom Eliseu	4.67	10.18	0.12
	C. do Araguaia	6.28	11.56	0.07
	Itaituba	7.76	14.35	-0.15
XGBoost	Mateiros	4.19	8.95	0.24
	Belém	8.11	12.95	0.17
	Dom Eliseu	4.98	9.78	0.19
	C. do Araguaia	7.32	11.22	0.12
	Itaituba	7.70	12.52	0.12
LightGBM	Mateiros	4.17	8.56	0.30
	Belém	7.97	13.35	0.17
	Dom Eliseu	4.68	10.18	0.22
	C. do Araguaia	6.64	11.56	0.18
	Itaituba	6.98	14.35	0.19

O segundo conjunto que além das variáveis base possuía as variáveis temporais derivadas também foram aplicados a treinamento e validação em todos os modelos, retornando os resultados conforme a Tabela 3.

Tabela 3. Desempenho dos modelos de previsão para precipitação com uso de variáveis derivadas

Algoritmo	Estação	MAE	RMSE	R^2
Regressão Linear	Mateiros	4.49	8.71	0.28
	Belém	8.22	13.02	0.11
	Dom Eliseu	5.45	9.68	0.21
	C. do Araguaia	6.56	11.20	0.23
	Itaituba	7.28	11.83	0.22
Random Forest	Mateiros	3.94	8.43	0.33
	Belém	8.27	12.97	0.12
	Dom Eliseu	4.43	9.69	0.21
	C. do Araguaia	6.00	11.28	0.12
	Itaituba	7.38	13.13	0.04
XGBoost	Mateiros	3.97	8.23	0.36
	Belém	8.22	12.71	0.15
	Dom Eliseu	4.86	9.66	0.22
	C. do Araguaia	6.89	11.06	0.15
	Itaituba	7.38	12.33	0.15
LightGBM	Mateiros	4.13	8.50	0.32
	Belém	7.96	12.56	0.18
	Dom Eliseu	4.93	9.68	0.21
	C. do Araguaia	6.38	10.90	0.18
	Itaituba	6.84	11.93	0.21

Posteriormente partimos a execução dos algoritmos para previsão da característica de temperatura máxima, aplicando o primeiro conjunto sem as variáveis temporais derivadas, e os resultados constam na Tabela 4

Tabela 4. Desempenho dos modelos de previsão para temperatura máxima sem uso de variáveis derivadas

Algoritmo	Estação	MAE	RMSE	R^2
Regressão Linear	Mateiros	0.94	1.19	0.75
	Belém	0.61	0.78	0.77
	Dom Eliseu	0.72	0.92	0.74
	C. do Araguaia	0.98	1.30	0.79
	Itaituba	0.59	0.73	0.90
Random Forest	Mateiros	0.97	1.20	0.74
	Belém	0.56	0.73	0.80
	Dom Eliseu	0.73	0.91	0.75
	C. do Araguaia	0.79	1.01	0.87
	Itaituba	0.59	0.74	0.90
XGBoost	Mateiros	1.02	1.25	0.72
	Belém	0.60	0.75	0.79
	Dom Eliseu	0.75	0.94	0.73
	C. do Araguaia	0.84	1.08	0.85
	Itaituba	0.64	0.81	0.88
LightGBM	Mateiros	0.96	1.22	0.73
	Belém	0.60	0.78	0.77
	Dom Eliseu	0.73	0.93	0.73
	C. do Araguaia	0.79	1.04	0.86
	Itaituba	0.61	0.77	0.89

Por fim executamos os modelos de previsão de temperatura máxima utilizando o conjunto de entrada que inclui as variáveis temporais derivadas, obtendo os resultados da Tabela 5.

Tabela 5. Desempenho dos modelos de previsão para temperatura máxima com uso de variáveis derivadas

Algoritmo	Estação	MAE	RMSE	R^2
Regressão Linear	Mateiros	0.80	1.01	0.82
	Belém	0.61	0.77	0.78
	Dom Eliseu	0.70	0.91	0.75
	C. do Araguaia	0.88	1.25	0.80
	Itaituba	0.55	0.69	0.91
Random Forest	Mateiros	0.68	0.88	0.86
	Belém	0.52	0.68	0.83
	Dom Eliseu	0.65	0.84	0.78
	C. do Araguaia	0.74	0.98	0.88
	Itaituba	0.51	0.63	0.93
XGBoost	Mateiros	0.70	0.90	0.86
	Belém	0.51	0.67	0.84
	Dom Eliseu	0.66	0.85	0.78
	C. do Araguaia	0.74	0.98	0.88
	Itaituba	0.55	0.66	0.92
LightGBM	Mateiros	0.73	0.95	0.84
	Belém	0.54	0.72	0.81
	Dom Eliseu	0.65	0.86	0.77
	C. do Araguaia	0.71	0.97	0.88
	Itaituba	0.52	0.66	0.92

4.1. Análise dos resultados

4.1.1. Impacto das Variáveis Temporais (Lags e Média Móvel)

Quando avaliamos os resultados das previsões de precipitação notamos uma melhoria modesta em MAE, RMSE e R^2 na maioria dos modelos. Por exemplo, em Mateiros, o *XGBoost* com variáveis temporais teve R^2 de 0.36 (contra 0.24 sem variáveis temporais), e em Itaituba, o *LightGBM* aumentou o R^2 de 0.19 para 0.21. Entretanto vemos exceções, como por exemplo em Belém que não houve um ganho em R^2 muito significativo, sugerindo que a precipitação para essa região pode ser menos dependente de padrões históricos e mais influenciada por fatores não capturados.

Importante ressaltar que a heterogeneidade geográfica e climática da região amazonica justifica a necessidade de modelos exclusivos para cada estação, buscando capturar padrões únicos de cada microclima. Notamos que estações que possuem padrões sazonais mais bem definidos se beneficiaram mais das variáveis temporais.

Ao avaliar os resultados de previsão para a variável de temperatura máxima nota-se que as melhoras nos valores das métricas são extremamente significativas quando são utilizadas as variáveis temporais. Em Mateiros, o R^2 do *Random Forest* teve um salvo de 0.74 para 0.86, e em Itaituba, o R^2 chegou a 0.93, indicando uma forte autocorrelação temporal na temperatura. É perceptível que existe mais consistência no impacto das variáveis

temporais já que todas as estações tiveram ganhos de desempenho, reforçando que a temperatura é mais previsível quando se inclui o histórico recente.

4.1.2. Comparação entre algoritmos

Ao avaliar a atuação dos algoritmos na previsão de precipitação os algoritmos mais estáveis, com menores MAE/RMSE são *XGBoost* e *LightGBM* para várias estações. O algoritmo de *Random Forest* apresentou problema de *overfitting* em Itaituba comn R^2 negativo (sem uso de variáveis temporais), o modelo memorizou ruído na fase de treinamento. Para regressão linear considerando apenas as métricas de desempenho nota-se um desempenho bastante competitivo comparado a algoritmos mais complexos, especialmente com uso de variáveis temporais.

Para as previsões de temperatura máxima algoritmos como *Random Forest* e *XGBoost* mostraram-se mais eficientes, com R^2 acima de 0.85 em várias estações. O *Random Forest* para estação de Itaituba atingiu um RMSE de 0.63, indicando alta precisão nas previsões. O *LightGBM* apresentou um desempenho levemente inferior ao *XGBoost* em alguns modelos. A regressão linear se mostou menos eficiente em comparação a outros algoritmos mas ainda sim com métricas de desempenho relevantes como para estação Mateiros com R^2 de 0.82, indicando que parte da variação de temperatura pode ser explicada de forma linear.

4.1.3. Comparação gráfica entre valor real e valor previsto

Além das métricas outra maneira de comprovar a eficiencia de um modelo de previsão é realizar uma análise entre valor previsto e valor real, para isso foi é utilizado os dados de validação que é um conjunto de dados não utilizados na etapa de treinamento. Como existem muitos modelos e variações o foco aqui será apenas em resultados expressivos para compreensão do funcionamento dos modelos.

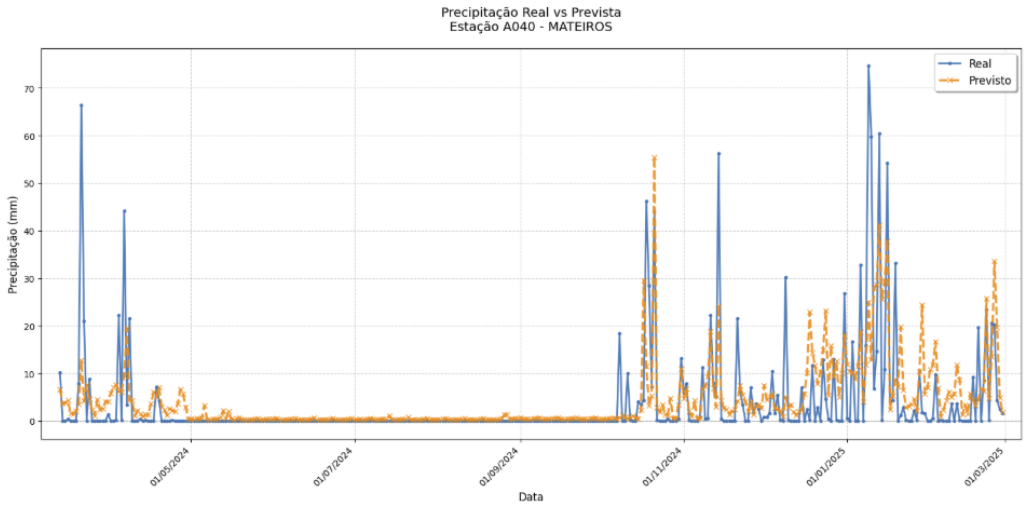

Figura 2. Modelo XGBoost para estação Mateiros com variáveis derivadas
Fonte: De autoria própria.

Na Figura 2 temos o modelo de previsão para precipitação que conseguiu maior valor na métrica R^2 que indica a porcentagem que o modelo é capaz de explicar a variação dos dados, para o caso da estação Mateiros usando *XGBoost* com uso de variáveis derivadas alcançou 0.36 indicando acerto em 36% das previsões, um valor que não é considerado ótimo mas ao avaliar o gráfico percebemos que o modelo foi capaz de perceber bem a sazonalidade entre épocas de maior frequencia de chuvas e épocas mais secas, mas tem dificuldade de prever valores de precipitação acima de 40 mm.

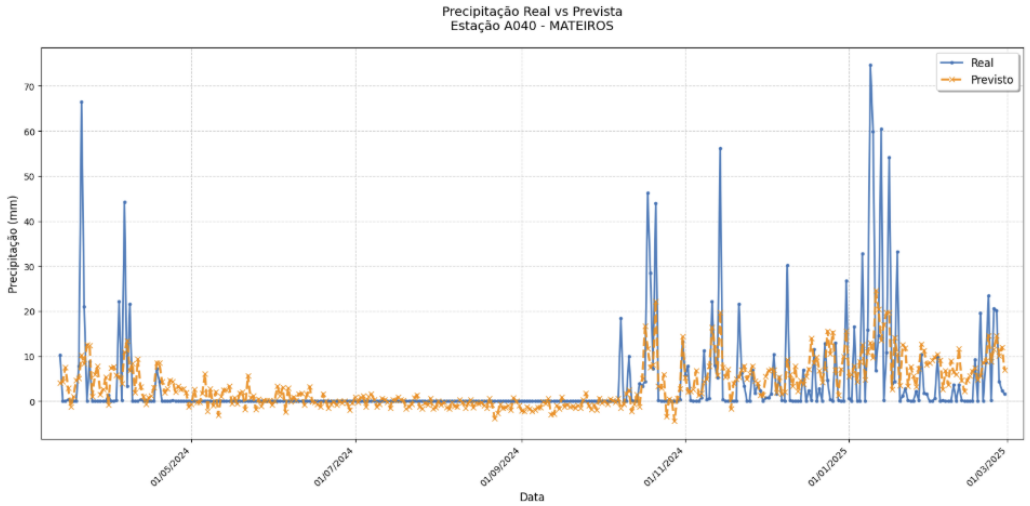

Figura 3. Modelo Regressão Linear para estação Mateiros com variáveis derivadas

Fonte: De autoria própria.

Na Figura 3 temos o modelo de previsão para precipitação que utiliza regressão linear para estação de Mateiros utilizando variáveis derivadas, o algoritmo de regressão linear conseguiu métricas competitivas em comparação com algoritmos mais complexos, com um R^2 para esse caso de 0.28, entretanto quando avaliamos o comportamento gráfico é perceptivel um problema na previsão de precipitação próxima de zero, pois o modelo prevê valores negativos para esta característica e isto não representa a realidade do cenário, indicando que o modelo não foi capaz de reconhecer o padrão de comportamento da variável alvo mesmo possuindo dados históricos para complementar as previsões.

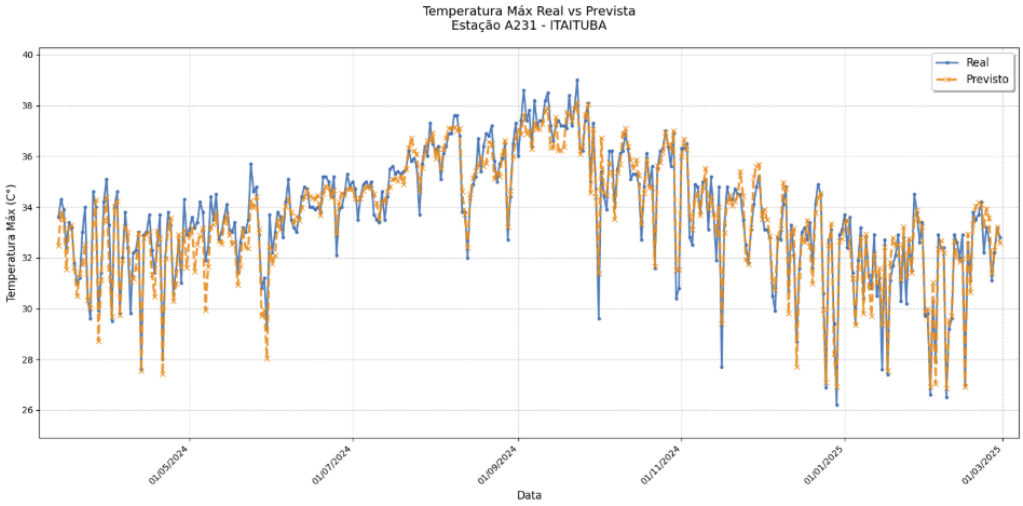

Figura 4. Modelo Random Forest para estação Itaituba com variáveis derivadas
Fonte: De autoria própria.

Na Figura 4 temos o modelo de previsão de temperatura máxima que obteve o melhor desempenho de R^2, utilizou o algoritmo de *Random Forest* com variáveis derivadas e alcançou valor de 0.93, indicando uma forte explicação pelo modelo aos dados previstos, é possível notar que todas as tendencias de de variação na métrica foram correspondidas nos valores previstos, com erros mínimos, mas fica claro que a estabilidade térmica da região garantiu ao modelo representações lineares claras, garantindo a alta performance.

5. Conclusão

Este estudo buscou avaliar a aplicabilidade de modelos de aprendizado de máquina na previsão de variáveis climáticas críticas para a Amazônia, região cujas especificidades microclimáticas e vulnerabilidades socioambientais demandam abordagens preditivas robustas e adaptadas. A análise comparativa entre algoritmos clássicos e ensembles, utilizando dados meteorológicos de cinco estações representativas, revelou informações importantes para o aprimoramento de sistemas de previsão na região.

Os resultados evidenciaram que a incorporação de variáveis temporais derivadas influenciou de maneira desigual as previsões, dependendo da natureza da variável climática analisada. Para fenômenos como precipitação, observou-se que a dependência de padrões históricos apresentou efeitos limitados em determinadas localidades, com ganhos de precisão variáveis entre os modelos. Esse comportamento sugere a atuação de fatores externos não mapeados — como alterações ambientais ou d inâmicas atmosféricas locais — que desafiam a capacidade preditiva dos algoritmos. Por outro lado, para variáveis relacionadas a temperaturas extremas, a utilização de dados históricos mostrou-se decisiva, com melhorias expressivas na acurácia dos modelos, indicando uma relação temporal mais marcante e previsível nesses casos.

Na análise comparativa entre técnicas de aprendizado de máquina, modelos baseados em estratégias ensemble demonstraram maior robustez em cenários de alta complexidade, especialmente para fenômenos com padrões não lineares. Em contrapartida, abordagens clássicas mantiveram relevância em contextos onde relações lineares predominam, reforçando a importância da escolha do algoritmo conforme as características

intrínsecas dos dados. A ocorrência de instabilidades em determinados cenários — como tendências de superajuste — destacou a necessidade de protocolos rigorosos de validação e personalização de parâmetros, especialmente em regiões com microclimas distintos, onde generalizações podem comprometer a confiabilidade das previsões.

A heterogeneidade geográfica da Amazônia, evidenciada pelas discrepâncias entre estações, reforça a importância de modelos customizados, capazes de incorporar variáveis locais e contextuais. Essa adaptação é crucial para aplicações práticas, como alertas precoces de secas ou inundações, que dependem não apenas da acurácia global, mas da confiabilidade em escalas menores.

Por fim, este trabalho contribui para a literatura ao propor um framework replicável de avaliação de modelos preditivos em contextos de alta variabilidade climática. Futuros estudos poderão expandir a análise usando correlações entre estações próximas expansão e generalização das previções, além de integrar dados socioambientais (ex.: cobertura vegetal, desmatamento, emissão de gases) para capturar interações entre clima e atividades humanas — um passo essencial para antecipar crises num cenário de mudanças aceleradas.

Referências

Artaxo, P., Rizzo, L. V., and Machado, L. A. T. (2024). Inteligência artificial e mudanças climáticas. *Revista USP*, (141):29–40.

Berk, R. A. (2008). *Statistical Learning from a Regression Perspective*. Springer Series in Statistics. Springer-Verlag, New York, NY, 1st edition.

Brownlee, J. (2019). Overfitting and underfitting with machine learning algorithms. Acesso em: 01 abr. 2025.

Faceli, K., Lorena, A. C., Gama, J., Almeida, T. A. d., and Carvalho, A. C. P. d. L. F. d. (2021). *Inteligência artificial: uma abordagem de aprendizado de máquina*. LTC.

Géron, A. (2021). *Mãos à Obra: Aprendizado de Máquina com Scikit-Learn & TensorFlow*. Alta Books, Rio de Janeiro, 1ª edition.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Series in Statistics. Springer, New York, NY, 2nd edition.

INMET (2025). Dados históricos de estações automáticas. Acesso em: 01 mar. 2025.

Lezaun, M. (2006). Que tiempo va hacer? *UNIÓN – Revista Iberoamericana de Educación Matemática*, 5:37–47.

Mitchell, T. M. (1997). *Machine Learning*. McGraw-Hill Series in Computer Science. McGraw-Hill Science/Engineering/Math, New York, 1st edition.

NASA (2024). Análise detalhada da NASA revela o futuro climático da terra e da amazônia. Acesso em: 1 mar. 2025.

Samuel, A. (1959). Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development*, 3:210–229.

Soto, T. (2013). Regression analysis. In Volkmar, F. R., editor, *Encyclopedia of Autism Spectrum Disorders*, pages 2538–8. Springer, New York, NY.

Strang, G. (2006). *Linear Algebra and Its Applications*. Cengage Learning, Boston, MA, 4th edition.