III SIMPOSIO BRASILEIRO DE BIOMETEOROLOGIA HUMANA
26 a 28 de agosto de 2026

# Previsão de asma em Salvador: uma abordagem com Machine Learning

**João Otavio Alves Accioly (1); Fabrício Daniel dos Santos Silva (2); Mário Henrique Guilherme dos Santos Vanderlei (3); Beatriz Jhúlia da Silva Barros Pereira (4); Rafaela Lisboa Costa (5); Heliofábio Barros Gomes (6); Matheus José de Arruda Lyra (7); Heitor Alves de Souza Santos (8);**
(1) Mestrando, Universidade Federal de Alagoas, joao.accioly@icat.ufal.br (2) Doutor, Universidade Federal de Alagoas, fabricio.santos@icat.ufal.br (3) Doutorando, Universidade Federal de Campina Grande, mario.vanderlei@icat.ufal.br (4) Graduanda, Universidade Federal de Alagoas, beatriz.pereira@icat.ufal.br (5) Doutora, Universidade Federal de Alagoas, rafaela.costa@icat.ufal.br (6) Doutor, Universidade Federal de Alagoas, heliofabio@icat.ufal.br (7) Doutor, Universidade Federal de Alagoas, matheus.lyra@icat.ufal.br (8) Mestrando, Universidade Federal de Alagoas, heitor.santos@icat.ufal.br

## RESUMO
Este trabalho avaliou a performance do modelo Random Forest na predição de casos de asma mensal em Salvador (2000-2020), em diferentes configurações de defasagens climáticas (lags 1-3 e 1-6 meses ) , testando o efeito da inclusão de uma variável temporal (tempo). Dados mensais de asma e de variáveis climáticas (precipitação, temperatura, vento, evapotranspiração, pressão, umidade, etc.) foram utilizados, tendo sido criadas defasagens de 1, 2, 3 e 6 meses. O modelo Random Forest foi utilizado com a estratégia de validação cruzada temporal (Time Series Split com 3 folds), tendo sido testadas quatro configurações principais. O modelo com a inclusão da variável tempo (M1) apresentou o melhor desempenho ($R^2$ = 0,92; RMSE = 29,6; MAE = 20,8), sendo que a variável tempo respondeu por 92,3% da importância preditiva, indicando uma expressiva tendência temporal dos casos na capital baiana. O modelo sem tempo (M2) também teve um bom desempenho ($R^2$ = 0,87; RMSE = 30,5; MAE = 21,4), demonstrando que somente com as variáveis climáticas em defasagem consegue-se incluir a maioria da variação sazonal da asma. Sem a variável tempo, a velocidade do vento em lag de 1 mês (VV_LAG1) foi o principal preditor climático (29,6%), seguido de VV_LAG2 (16,3%) e TMIN_LAG3 (11,9%), sugerindo que as condições de vento e temperatura mínima do mês anterior impactaram bastante os casos de crises asmáticas em Salvador.Os modelos com defasagens estendidas até 6 meses (M3 e M4) apresentaram desempenho inferior ($R^2$ = 0,91 e 0,85, respectivamente), sugerindo que lags mais longas não trouxeram ganhos adicionais em sua capacidade preditiva. Portanto, a inclusão da variável temporal reflete uma maior acurácia, mas rendimentos mais elevados podem ainda ser explorados por modelos baseados exclusivamente nas variáveis climáticas defasadas como uma alternativa para estudos voltados em estudar como os fatores ambientais afetam a asma em Salvador sem o efeito mascarador da tendência temporal.
Palavras-chave: RandomForest; Clima: Defasagem Temporal.

## 1. INTRODUÇÃO

A asma pode ser definida como uma doença respiratória crônica, afetando 339 milhões de pessoas em todo o mundo e apresentando um relevante problema de saúde pública (XIONG et al., 2023). Em 2019, 461 mil mortes relacionadas ao quadro de asma ocorreram, o que denota a gravidade da doença (ALKOBAISI et al., 2025). Estudos epidemiológicos têm demonstrado que os fatores ambientais, especialmente variáveis climáticas e poluentes atmosféricos, exercem importância na ocorrência e na gravidade das crises asmáticas (REIS et al., 2025). Nos últimos anos, abordagens de machine learning têm sido cada vez mais utilizadas na área da saúde, pois são capazes de lidar com relações não-lineares e interações complexas entre variáveis (XIONG et al., 2023).

O Random Forest é um método de aprendizado de conjunto que constrói um conjunto de árvores de decisão e combina suas previsões, apresentando robustez contra o overfitting e fornecendo notas de importância das variáveis preditoras (ALKOBAISI et al., 2025). Estudos comparativos têm mostrado que o Random Forest apresenta um desempenho superior a outros algoritmos em relação à predição de risco para a asma (DAELI & SANJAYA, 2025; AMINU & YUSUF, 2026).

A introdução de defasagens temporais (lags) das variáveis climáticas capacita a captura de efeitos de exposição no passado na expressão dos sintomas asmáticos (REIS et al., 2025). Pesquisas em diferentes ambientes geográficos indicam que as hospitalizações por asma apresentam diferenças

espaço-temporalmente significativas, acompanhadas de padrões sazonais distintos que estão relacionados com o clima local (WANG et al., 2026; SIMUNOVIC et al., 2021). Nesse sentido, este estudo analisa a habilidade preditiva do modelo Random Forest em estimar os casos mensais de asma em Salvador, comparando diferentes opções de defasagens climáticas e avaliando o impacto da inclusão da variável tempo no desempenho do modelo.

## 2. OBJETIVO

Avaliar a eficácia do modelo Random Forest na previsão de casos mensais de asma em Salvador por meio da comparação, entre diversas configurações de lags climáticos, de lags de 1 a 3 meses e de lags de 1 a 6 meses, e verificar a influência da inclusão da variável tempo sobre o desempenho preditivo do modelo.

## 3. MÉTODO

### 3.1. Dados e pré-processamento

Foram utilizados dados mensais de atendimento a casos de asma, em Salvador, entre 2000 e 2020. As variáveis climáticas englobaram: precipitação (PREC), temperatura média (TMED), temperatura mínima (TMIN), a velocidade do vento (VV) , evapotranspiração (EVP), deficit de pressão de vapor (DCP), insolação (INS), pressão atmosférica (PRES) , umidade relativa do ar (UR) e outras variáveis derivadas. Para captar os efeitos retardados das condições meteorológicas na ocorrência da asma, foram criadas defasagens (lags) de 1, 2, 3 e 6 meses para todas as variáveis climáticas, resultando em features como VV_LAG1 , TMIN_LAG3 , EVP_LAG6, etc. (REIS et al. 2025). Adicionalmente, para algumas configurações, foi incluída uma variável tempo, representando uma tendência linear sequencial (1, 2, 3,..., N), que teve por finalidade captar padrões grande de longo prazo que não eram capturados pelas variáveis climáticas (WANG et al. 2026).

### 3.2. Modelagem e validação

Utilizou-se o algoritmo Random Forest Regressor com os seguintes hiperparâmetros: n_estimators = 500, max_depth = 8, min_samples_split = 5 e min_samples_leaf = 3 (BREIMAN, 2001; REIS et al., 2025). As variáveis foram padronizadas com StandardScaler a cada fold, visando evitar vazamento de dados. A validação ocorreu através de validação cruzada temporal (Time Series Split com 3 folds), respeitando rigidamente a ordem das observações, com tamanho de teste fixado em aproximadamente 15% da série (XIONG et al., 2023).

### 3.3. Configurações Experimentais

Cinco configurações diferentes foram testadas, de acordo com o que é recomendado na literatura para estudos de predição de asma com variáveis ambientais (REIS et al., 2025; SIMUNOVIC et al., 2021): M1: Lags 1, 2 e 3 com variável tempo; M2: Lags 1, 2 e 3 sem variável tempo; M3: Lags 1, 2, 3 e 6 com variável tempo; M4: Lags 1, 2, 3 e 6 sem variável tempo.

### 3.4. Métricas e análise de importância

O desempenho dos modelos foi avaliado pelas métricas: coeficiente de determinação ($R^2$), raiz do erro quadrático médio (RMSE) e erro absoluto médio (MAE) (XIONG et al., 2023; ALKOBAISI et al., 2025). A importância de cada variável preditora foi extraída diretamente do modelo Random Forest através do atributo feature_importances_, que quantifica a média da redução da impureza (Mean Decrease in Impurity) de cada feature (BREIMAN, 2001; AMINU & YUSUF, 2026).

## 4. RESULTADOS

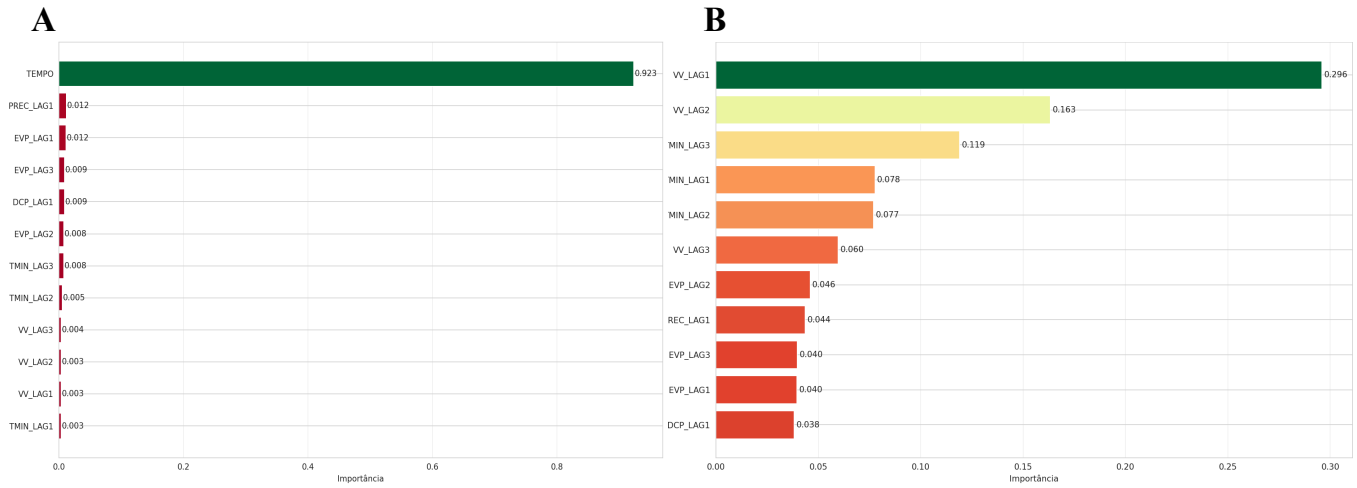
A Tabela 1 apresenta o desempenho obtido pelos modelos Random Forest nas diferentes configurações testadas. O modelo M1 (lags 1,2,3 + tempo) teve a melhor performance dentre todas as configurações, com $R^2 = 0,92$, RMSE = 29,6 e MAE = 20,8, o que significa que 92% da variação dos casos de asma é explicada pelo modelo quando a variável temporal é incluída. O modelo M2 (lags 1,2,3 - tempo) também obteve bom desempenho, com $R^2 = 0,87$, RMSE = 30,5 e MAE = 21,4, evidenciando que apenas os lags das variáveis climáticas conseguem explicar boa parte da variabilidade sazonal da enfermidade, mesmo sem a presença da tendência temporal. Os modelos com lags estendidos até 6 meses (M3 e M4) tiveram desempenho ligeiramente inferior, com $R^2 = 0,91$ e 0,85 respectivamente, sugerindo que a adição de lags mais longos não trouxe ganho adicional ao poder preditivo.

Tabela 1 - Desempenho dos experimentos modelo *random forest.* (Autores, 2026)

| Modelo | Lags | Tempo | R^2 | RMSE | MAE |
|---|---|---|---|---|---|
| M1 | 1,2,3 | Sim | 0,92 | 29,6 | 20,8 |
| M2 | 1,2,3 | Não | 0,87 | 30,5 | 21,4 |
| M3 | 1,2,3,6 | Sim | 0,91 | 31,2 | 22,1 |
| M4 | 1,2,3,6 | Não | 0,85 | 32,8 | 23,5 |

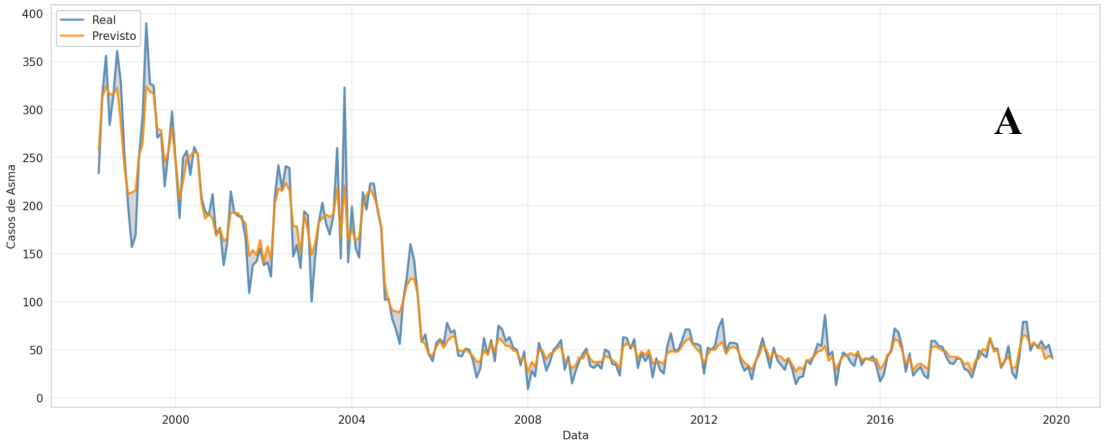
A análise da importância das variáveis (Figura 2) revelou que, no M1, a variável tempo respondeu por 92,3% da importância preditiva, indicando uma forte tendência temporal dos casos. As variáveis climáticas restantes apresentaram contribuições inferiores a 1%, com destaque para PREC_LAG1, EVP_LAG1 e DCP_LAG1, todas com importância < 1,2%. Sem a variável tempo inserida (Modelo M2), o quadro se inverteu: a velocidade do vento com defasagem de um mês (VV_LAG1) foi o preditor que obteve a mais elevada importância (29,6%), seguida pela VV_LAG2 (16,3%) e TMIN_LAG3 (11,9%). A temperatura mínima (TMIN) apresentou-se em defasagens 1, 2 e 3 meses, totalizando importância significativa.

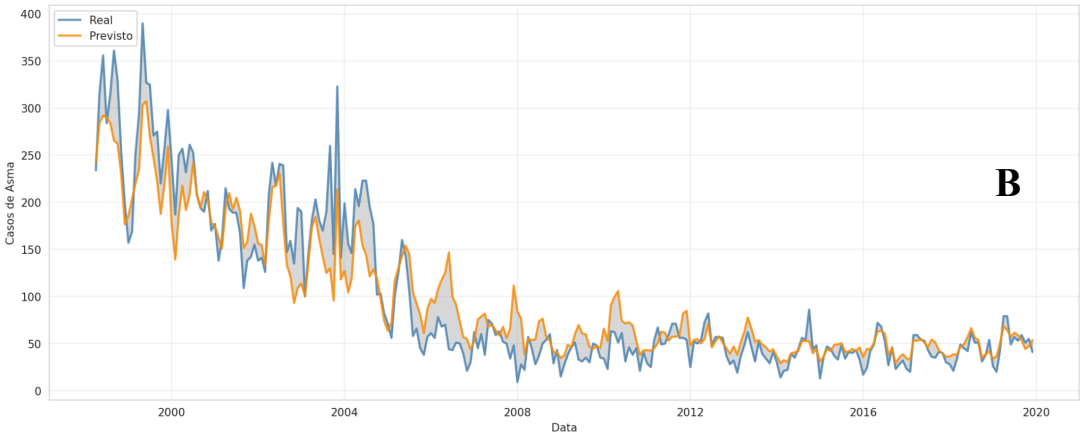
Figura 2 - Importância das variáveis preditoras variando o parâmetro tempo.




As Figuras 3A e 3B exibem as séries históricas dos valores observados e os previstos para os modelos com e sem a variável tempo. O modelo com tempo (Figura 3A, M1) apresentou um ajuste muito bom aos dados observados, reproduzindo bem os picos sazonais e a tendência de longo prazo da série histórica da variável (2000-2020). O modelo sem tempo (Figura 3B, M2), mesmo com pequenos erros nos picos extremos, conseguiu captar as variações sazonais gerais dos casos de asma. Isso sugere que apenas as variáveis climáticas defasadas já são capazes de explicar grande parte da variação observada da doença. Em suma, o M1 apresenta melhor desempenho visual, mas o M2 parece uma alternativa viável para estudos com foco nos fatores climáticos, sem o efeito mascarador da tendência temporal.

Figura 3 - Série temporal de previsão vs observado para os experimentos M1 e M2.






## 5. CONCLUSÕES

O modelo Random Forest, que inclui a variável temporal (M1), foi o melhor na predição de casos de asma em Salvador ($R^2$ = 0,92; RMSE = 29,6; MAE = 20,8), com a variável tempo explicando 92,3% de importância preditiva. Essa resultado evidencia a forte tendência temporal dos casos de asma na capital baiana, sugerindo que fatores não climáticos, como crescimento populacional, maior acurácia do diagnóstico ou mudança um padrão de atendimento ao longo dos anos, têm forte atuação na determinação da doença. O modelo sem tempo (M2), apresentou bom resultado também, ($R^2$ = 0,87; RMSE = 30,5; MAE = 21,4), mostrando que apenas as variáveis climáticas defasadas conseguem explicar boa parte da variabilidade sazonal da asma em Salvador. Na ausência da variável temporal, a velocidade do vento com defasagem de um mês (VV_LAG1) foi o principal preditor climático (29,6%), seguido por VV_LAG2 (16,3%) e TMIN_LAG3 (11,9%), indicando que condições de vento e temperatura mínima no mês anterior têm forte aclamação nas crises asmáticas do município. Os modelos com defasagens maiores que 1, em até 6 meses (M3 e M4), embora tivessem desempenho bem inferior ($R^2$ = 0,91 e 0,85, respectivamente), tiveram pior do que o M1 e M2, indicando que a inclusão de lags mais longos não trouxe ganhos adicionais na predição do modelo.

## REFERÊNCIAS BIBLIOGRÁFICAS

ALKOBAISI, S. et al. A hybrid approach for forecasting peak expiratory flow rate in asthma patients using combined linear regression and random forest model. **PLoS One**, v. 20, n. 8, p. e0326036, 2025.

AMINU, I.; YUSUF, M. M. Developing Enhanced Asthma Severity Predictive Model using Random Forest Algorithm. **Journal of Science Innovation & Technology Research**, v. 11, n. 9, 2026.

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

DAELI, B. B. F.; SANJAYA, U. P. Comparative Analysis of Logistic Regression, Random Forest, and SVM for Asthma Risk Prediction Using Demographic, Clinical, and Environmental Features. **Journal of Applied Informatics and Computing**, v. 9, n. 5, p. 2187-2194, 2025.

REIS, J. S. dos et al. Predicting asthma hospitalizations from climate and air pollution data: a machine learning-based approach. **Climate**, v. 13, n. 2, p. 23, 2025.

SIMUNOVIC, M. et al. Spatiotemporal characteristics of asthma emergency department presentations in diverse geographical and climatic regions, Queensland, Australia. **Emergency Medicine Australasia**, v. 33, n. 4, p. 623-630, 2021.

WANG, J. et al. Spatial and temporal Patterns in Childhood and Adolescent Asthma Hospitalisations in Queensland, Australia: A 20‐Year Ecological Study Across Climate Zones. **Medical Journal of Australia,** v. 224, n. 2, p. e70145, 2026.

XIONG, S. et al. Machine learning for prediction of asthma exacerbations among asthmatic patients: a systematic review and meta-analysis. **BMC pulmonary medicine**, v. 23, n. 1, p. 278, 2023.

## AGRADECIMENTOS

O autor agradece ao CNPq pela bolsa de mestrado concedida (processo nº 130879/2026-6) referente a chamada pública CNPq Nº 12/2025 - PROGRAMA INSTITUCIONAL DE BOLSAS DE PÓS-GRADUAÇÃO (PIBPG) - CICLO 2026.