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

Interpretabilidade da predição de asma em Salvador: aplicação de Random Forest e SHAP Analysis

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); (9) Dayane Alexandre Correia

(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 (9) Doutoranda, Universidade Federal de Alagoas, dayane.correia@fau.ufal.br

RESUMO

Este estudo utilizou o modelo Random Forest incluindo variáveis climáticas de defasagens (*lags*) de 1, 2 e 3 meses para predição dos casos mensais da asma em Salvador, entre 2000 e 2020, utilizando a interpretação da análise SHAP para capturar a importância e a direção dos efeitos da variável preditora, excluindo a variável de tempo para que essa influência pudesse ser isolada da exclusiva influência de fatores ambientais. O modelo alcançou R^2 = 0,87, RMSE = 30,5 e MAE = 21,4, indicando que as variáveis climáticas de LAG de defasagem explicaram 87 % da variação sazonal da doença. A análise SHAP indicou que a velocidade do vento defasada em um mês (VV_LAGI) foi a preditora mais importante, com média da contribuição de 31,0, seguida por VV_LAG2 e TMIN_LAG3, com 18,0 e 16,0, respectivamente. O summary plot SHAP sugere que a velocidade do vento mais alta está positivamente associada ao aumento da predição, provavelmente pela dispersão dos alergênios, ao passo que a menor temperatura mínima, especialmente com defasagem de 3 meses, poderia estar contribuindo positivamente para aumento dos casos, corroborando associação entre ar frio e exacerbações asmáticas. Como limitação, o estudo não contemplou os dados de poluentes atmosféricos e dados socioeconômicos.Conclui-se que o *Random Forest* aliado ao SHAP é uma ferramenta eficaz para predição e interpretação da asma em Salvador, recomendando-se para pesquisas futuras a inclusão de poluentes e a comparação com outros algoritmos.

Palavras-chave: Asma; Random Forest; SHAP; Defasagem temporal.

1. INTRODUÇÃO

As condições do ambiente exercem um impacto direto na saúde respiratória, sobretudo em áreas urbanas onde as condições de qualidade do ar e as variáveis meteorológicas têm ampla variabilidade ao longo do tempo. A asma, uma das síndromes crônicas que afeta milhões de pessoas em todo o mundo, frequentemente tem seu surgimento atraído e associado a fatores climáticos tais como temperatura, umidade, precipitação e, principalmente, à velocidade do vento, que atua no transporte de alérgenos e poluentes atmosféricos (SIMUNOVIC et al., 2021; REIS et al., 2025).

Nesse sentido, técnicas de machine learning têm se mostrado como ferramentas poderosas para o desenvolvimento de modelos de fenômenos complexos na saúde. Em contraste aos métodos estatísticos tradicionais, algoritmos como o Random Forest se destacam pela capacidade de captar relações não-lineares e interações entre múltiplas variáveis, sem a necessidade de pressupostos distribucionais fortes (XIONG et al., 2023). Um dos principais desafios na modelagem de doenças respiratórias é a incorporação dos efeitos retardados das condições meteorológicas. As defasagens temporais (lags) possibilitam representar a influência da exposição anterior sobre a manifestação atual dos sintomas, permitindo reconhecer que os efeitos do clima sobre a saúde não se dão de maneira imediata, podendo levar semanas para que sejam expressados (REIS et al., 2025).

Além de prever, a interpretabilidade dos modelos de machine learning tem recebido cada vez mais destaque na literatura. Ferramentas como SHAP (SHapley Additive exPlanations) possibilitam não apenas saber quais as variáveis mais relevantes, mas também entender a direção do efeito delas sobre a predição (LUNDBERG & LEE, 2017).

Diante do exposto, este estudo tem por proposta a aplicação do modelo Random Forest com variáveis climáticas defasadas (lag de 1, 2 e 3 meses) para predição dos casos mensais de asma em Salvador, usando a análise SHAP para interpretar a relevância e o efeito direcional de cada preditor.

2. OBJETIVO

Interpretar a predição dos casos mensais de asma em Salvador através do modelo Random Forest, ligado a SHAP, verificando quais variáveis climáticas desfazadas (lags de 1, 2 e 3 meses) são as mais influentes e o sentido de seus efeitos sobre os casos da doença, excluindo a variável de tempo tempo para filtrar a função específica das variáveis climáticas deixando intocados os efeitos dos fatores climáticos.

3. MÉTODO

3.1. Dados e pré-processamento

Dados mensais dos casos de asma em Salvador, no período de 2000 a 2020, foram utilizados. As variáveis climáticas foram: precipitação (PREC), temperatura mínima (TMIN), velocidade do vento (W), evapotranspiração (EVP) e o déficit de pressão de vapor (DCP). Para realizar a captura de efeitos retardados das condições meteorológicas, todas as variáveis foram defasadas (lags) em 1, 2 e 3 meses, resultando em features, por exemplo, como VV_LAG1, TMIN_LAG3, EVP_lag2, entre outras (REIS et al., 2025). A variável temporal tempo não foi considerada neste experimento.

3.2. Modelagem e validação

O algoritmo Random Forest Regressor foi usado, com os parâmetros hiper: n_estimators = 500, max_depth = 8, min_samples_split = 5 e min_samples_leaf = 3 (BREIMAN, 2001). As variáveis foram padronizadas pelo StandardScaler. A validação foi feita via validação cruzada temporal (Time Series Split com 3 folds), mantendo a ordem cronológica das observações, onde o tamanho do teste foi fixado em cerca de 15% da série (XIONG et al., 2023).

3.1. Métricas e Análise SHAP

Os SHAP value para a variável i em uma dada instância são a média ponderada das contribuições marginais desta variável quando outras são adicionadas em todas as possíveis coalizões, sendo calculados pela equação:

$$\phi_i = \Sigma_{S \subseteq F \setminus \{i\}} \frac{|S|!*(|F|-|S|-1)!}{|F|!} [f(S \cup \{i\} - f(S)] \quad \text{Equação 1}$$

em que F é o conjunto de todas as variáveis preditoras, S é um subconjunto de variáveis que não contém a variável i, |S| é o número de variáveis no subconjunto S, f(S) é a predição do modelo considerando as variáveis do subconjunto S, e a diferença $f(S \cup \{i\}) - f(S)$ mostra a contribuição marginal associada à variável i ao ser acrescentada à coalizão S;

O SHAP está dentro de um modelo de explicação aditiva linear, onde o valor base é a expectativa da predição dada a distribuição dos dados, e cada variável é acompanhada de seu respectivo SHAP value e de um indicador binário da sua presença. Uma propriedade do SHAP é a eficiência, que garante que a soma dos SHAP values de todas as variáveis é igual à diferença entre a predição atual e o valor base. Para modelos de caixa-preta, utiliza-se o Kernel SHAP, que resolve um problema de minimização ponderada cuja função peso depende do tamanho da coalizão e do número total de variáveis.

A importância global de cada variável é dada pela média dos valores absolutos de SHAP em todas as instâncias:

$$Importância_i = \frac{1}{n} \sum_{j=1}^{n} |\phi_i^{(j)}| \quad \text{Equação 2}$$

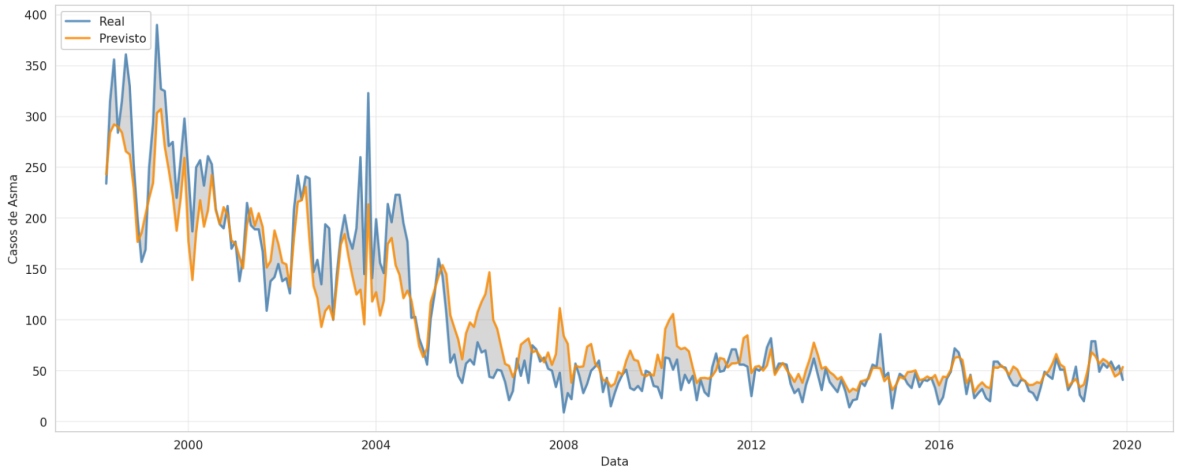
onde n é o número total de instâncias e $\phi_i^{(j)}$ é o SHAP value da variável i na instância j.

4. RESULTADOS

O modelo Random Forest, incluindo variáveis climáticas lags (defasadas 1, 2 e 3), teve um desempenho preditivo satisfatório, com o coeficiente de determinação $R^2 = 0,87$, um erro médio quadrático (RMSE) de 30,5 e um erro absoluto médio (MAE) de 21,4. Esses números indicam que as variáveis climáticas lag são capazes de explicar 87% da variação dos casos mensais de asma em Salvador. A Figura 1 mostra a série temporal dos valores observados e previstos pelo modelo ao longo do período de 2000 a 2020.

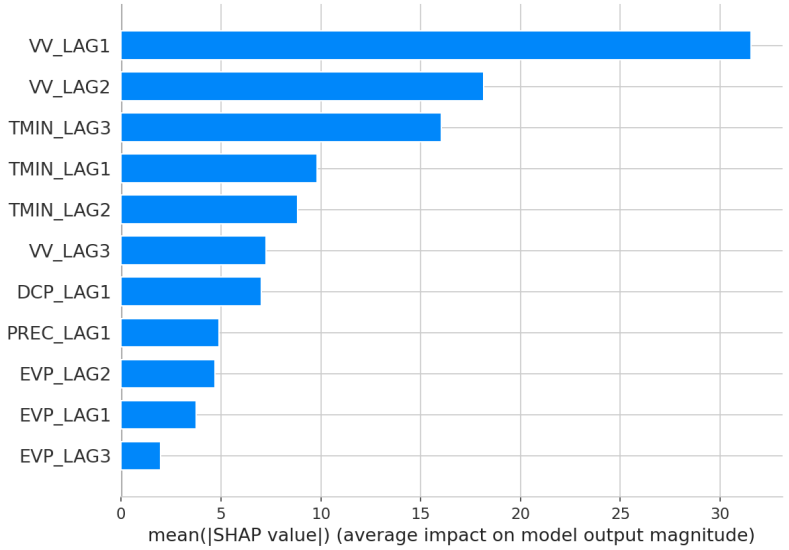
Percebe-se que o modelo consegue reproduzir bem as oscilações sazonais dos casos de asma, captando os picos e vales da série histórica. Apesar de pequenos desvios nos picos extremos, a adequação visual demonstra que as variáveis climáticas lags conseguem explicar boa parte da variabilidade temporal da referida doença.

Figura 1 - Série temporal de previsão vs observado.



A Figura 2 revela a importância global das variáveis preditoras em função dos valores médios absolutos de SHAP. A variável VV_LAG1 (velocidade do vento com defasagem de 1 mês) obteve a maior importância (31,0), seguida de VV_LAG2 (18,0) e TMIN_LAG3 (16,0). As variáveis de temperatura mínima se destacaram nas três defasagens (TMIN_LAG1 com 9,0; TMIN_LAG2 com 8,0; TMIN_LAG3 com 16,0), apresentando assim uma relevância estável no tempo. Variáveis como DCP_LAG1 (6,0), PREC_LAG1 (5,0) e as defasagens das variáveis de evapotranspiração (EVP_LAG2 com 4,5; EVP_LAG1 com 3,5; EVP_LAG3 com 2,5) obtiveram relevância secundária.

Figura 2 - Importância das features via SHAP

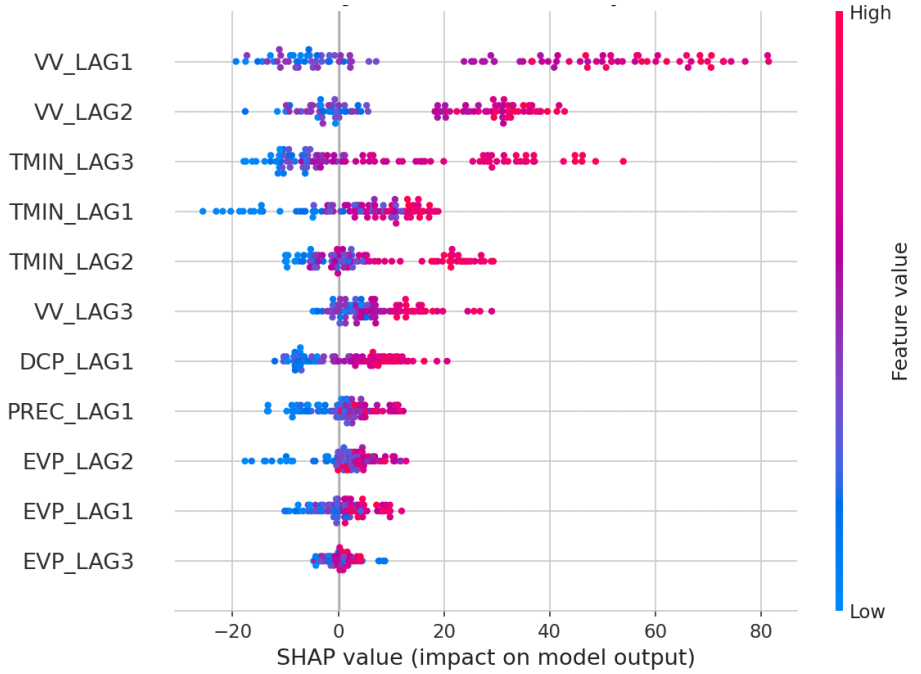


A Figura 3 mostra o gráfico summary plot SHAP, que combina a importância das variáveis com a direção do seu efeito na predição. Para a variável VV_LAG1 (velocidade do vento com lag de 1 mês), é possível observar que valores maiores (pontos vermelhos) estão associados a efeitos SHAP positivos, ou seja, maiores velocidades do vento do mês anterior estão relacionadas ao aumento da predição de casos de asma. Para VV_LAG2, observa-se um padrão semelhante, mas em menor magnitude. Para TMIN_LAG3 (temperatura mínima com lag de 3 meses), valores mais baixos (pontos azuis) normalmente tendem a gerar efeitos SHAP positivos, o que indica que temperaturas mínimas mais baixas com lag de 3 meses estão associadas ao aumento de casos.

As variáveis de temperatura mínima nas defasagens restantes (LAG1, LAG2) têm padrão similar, onde os valores mais baixos contribuem positivamente para a predição. Para as variáveis de

evapotranspiração (EVP_LAG1, EVP_LAG2 e EVP_LAG3), observa-se que valores maiores estão associados a efeitos positivos, mas com contribuições de menor magnitude.

Figura 3 - Summary plot SHAP para as principais variáveis preditoras



5. CONCLUSÕES

O modelo Random Forest, juntamente com a análise SHAP, alcançou o intuito de interpretar a predição mensal da asma em Salvador, a qual indicou a velocidade do vento com um mês de defasagem (VV_LAG1) como a variável mais importante (SHAP = 31,0), seguida de VV_LAG2 (18,0) e TMIN_LAG3 (16,0). A análise direcional demonstrou que os maiores valores de vento e os menores valores de temperatura mínima estão associados ao aumento das predições, corroborando a ação de fatores climáticos sobre a ocorrência da doença.

O modelo apresentou desempenho satisfatório (R^2 = 0,87), sugerindo que variáveis ambientais em defasagem podem explicar boa parte da variabilidade sazonal da asma em Salvador. Como limitação, o estudo não considerou dados de poluentes atmosféricos (material particulado, ozônio, dióxido de nitrogênio) e dados socioeconômicos e de qualidade da habitação, fatores que podem interagir com variáveis climáticas e afetar as ocorrências de asma. Sugere-se a inclusão de dados de poluentes atmosféricos e outros algoritmos de aprendizado de máquina (XGBoost, SVM) na próxima abordagem para desenvolver uma comparação de desempenho.

REFERÊNCIAS BIBLIOGRÁFICAS

BREIMAN, Leo. Random Forests. **Machine Learning,** v. 45, n. 1, p. 5-32, 2001.

LUNDBERG, Scott M.; LEE, Su-In. A unified approach to interpreting model predictions. **Advances in neural information processing systems**, v. 30, 2017.

REIS, Jean Souza 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, Marko 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.

XIONG, Shiqiu 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.