














## Article

# Predicting Asthma Hospitalizations from Climate and Air Pollution Data: A Machine Learning-Based Approach

Jean Souza dos Reis <sup>1,2</sup> , Rafaela Lisboa Costa <sup>2</sup> , Fabricio Daniel dos Santos Silva <sup>2,\*</sup> , Ediclê Duarte Fernandes de Souza <sup>3</sup> , Taisa Rodrigues Cortes <sup>1</sup> , Rachel Helena Coelho <sup>1</sup> , Sofia Rafaela Maito Velasco <sup>1</sup>, Danielson Jorge Delgado Neves <sup>1</sup>, José Firmino Sousa Filho <sup>1</sup> , Cairo Eduardo Carvalho Barreto <sup>4</sup> , Jório Bezerra Cabral Júnior <sup>5</sup> , Herald Souza dos Reis <sup>6</sup> , Keila Rêgo Mendes <sup>7</sup>, Mayara Christine Correia Lins <sup>2</sup>, Thomás Rocha Ferreira <sup>2</sup>, Mário Henrique Guilherme dos Santos Vanderlei <sup>2</sup>, Marcelo Felix Alonso <sup>8</sup>, Glauber Lopes Mariano <sup>2</sup> , Heliofábio Barros Gomes <sup>2</sup>  and Helber Barros Gomes <sup>2</sup> 

- <sup>1</sup> Centro de Integração de Dados e Conhecimentos para Saúde, Instituto Gonçalo Moniz, Fundação Oswaldo Cruz, Salvador 41745-715, Brazil; jean.reis@fiocruz.br (J.S.d.R.); taisa.cortes@fiocruz.br (T.R.C.); rachel.coelho@fiocruz.br (R.H.C.); sofia.velasco@fiocruz.br (S.R.M.V.); danielson.neves@fiocruz.br (D.J.D.N.); jose.sousa@ufba.br (J.F.S.F.)
- <sup>2</sup> Instituto de Ciências Atmosféricas, Universidade Federal de Alagoas, Maceió 57072-900, Brazil; rafaela.costa@icat.ufal.br (R.L.C.); mayara.lins@icat.ufal.br (M.C.C.L.); thomas.ferreira@icat.ufal.br (T.R.F.); mario.vanderlei@icat.ufal.br (M.H.G.d.S.V.); glauber.mariano@icat.ufal.br (G.L.M.); heliofabio@icat.ufal.br (H.B.G.); helber.gomes@icat.ufal.br (H.B.G.)
- <sup>3</sup> Center for Sci-Tech Research in Earth System and Energy-CREATE, Instituto de Investigação e Formação Avançada-IIFA, Earth Remote Sensing Laboratory (EaRS Lab), University of Évora, 7000-671 Évora, Portugal; edicle.duarte@uevora.pt
- <sup>4</sup> Coordenação de Hidrologia, Centro Gestor e Operacional do Sistema de Proteção da Amazônia (Censipam), Belém 66617-420, Brazil; cairo.barreto@sipam.gov.br
- <sup>5</sup> Instituto de Geografia, Desenvolvimento e Meio Ambiente, Universidade Federal de Alagoas, Maceió 57072-900, Brazil; jorio.cabral@igdema.ufal.br
- <sup>6</sup> Laboratory of Sexually Transmitted Infections, Bacteriology and Mycology Section, Evandro Chagas Institute (IEC), Ananindeua 67030-000, Brazil; heraldris@iec.gov.br
- <sup>7</sup> Departamento de Ciências Climáticas e Atmosféricas, Universidade Federal do Rio Grande do Norte, Natal 59078-970, Brazil; keila.mendes.859@ufrn.edu.br
- <sup>8</sup> Faculdade de Meteorologia, Universidade Federal de Pelotas, Pelotas 96010-610, Brazil; marcelo.alonso@ufpel.edu.br
- \* Correspondence: fabricio.santos@icat.ufal.br



Academic Editors: Roberto Ariel Abeldaño Zuñiga and Gabriela Narcizo de Lima

Received: 27 November 2024

Revised: 20 January 2025

Accepted: 21 January 2025

Published: 24 January 2025

**Citation:** Reis, J.S.d.; Costa, R.L.; Silva, F.D.S.; de Souza, E.D.F.; Cortes, T.R.; Coelho, R.H.; Velasco, S.R.M.; Neves, D.J.D.; Sousa Filho, J.F.; Barreto, C.E.C.; et al. Predicting Asthma Hospitalizations from Climate and Air Pollution Data: A Machine Learning-Based Approach. *Climate* **2025**, *13*, 23. <https://doi.org/10.3390/cli13020023>

**Copyright:** © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

**Abstract:** This study explores the predictability of monthly asthma notifications using models built from different machine learning techniques in Maceió, a municipality with a tropical climate located in the northeast of Brazil. Two sets of predictors were combined and tested, the first containing meteorological variables and pollutants, called exp1, and the second only meteorological variables, called exp2. For both experiments, tests were also carried out incorporating lagged information from the time series of asthma records. The models were trained on 80% of the data and validated on the remaining 20%. Among the five methods evaluated—random forest (RF), eXtreme Gradient Boosting (XGBoost), Multiple Linear Regression (MLR), support vector machine (SVM), and K-nearest neighbors (KNN)—the RF models showed superior performance, notably those of exp1 when incorporating lagged asthma notifications as an additional predictor. Minimum temperature and sulfur dioxide emerged as key variables, probably due to their associations with respiratory health and pollution levels, emphasizing their role in asthma exacerbation. The autocorrelation of the residuals was assessed due to the inclusion of lagged variables in some experiments. The results highlight the importance of pollutant and meteorological factors in predicting asthma cases, with implications for public health monitoring. Despite the limitations presented and discussed, this study demonstrates that forecast accuracy improves when a wider range of lagged variables are used, and indicates the suitability of RF for health datasets with complex time series.

**Keywords:** health data analysis; epidemiology; respiratory diseases; predictive modeling

---

## 1. Introduction

Asthma is a chronic health condition that affects all age groups globally, and is characterized by a heterogeneous inflammatory disease of the lungs, with variable symptoms such as coughing, wheezing, shortness of breath, chest tightness, and variable airflow limitation [1]. It is estimated that around 339 million people worldwide currently live with this condition, with many experiencing seasonal and daily variations in their symptoms, with different levels of severity throughout their lives [2]. Although most individuals manage to control their symptoms, there is a constant risk of acute crises that can lead to hospitalization and, in severe cases, death [3,4]. Over the last 50 years, there has been a significant increase in the prevalence of asthma, especially in children in industrialized countries, largely due to growing environmental pollution and exposure to irritants [5]. This reality makes it essential to study asthma in depth, not only in terms of its prevalence but also its mortality, in order to guide public health policies and mitigation strategies.

Asthma not only represents a significant public health challenge, but is also associated with alarming mortality rates. In 2019, approximately 461,000 people died from asthma-related complications worldwide, which demonstrates the severity of the condition [6]. In addition to its prevalence, which affects around 262 million individuals, asthma is a chronic disease that profoundly impacts patients' quality of life, as it currently has no definitive cure [7]. Epidemiological studies have shown a strong association between exposure to air pollutants and the worsening of asthma, suggesting that environmental factors play a crucial role in the progression of the disease [8]. The literature also points out that elements such as air pollution, tobacco smoke, climatic conditions, allergens such as pollen, and pathogens such as influenza viruses are among the main factors that exacerbate asthma [6]. In the face of climate change, the situation becomes even more critical, since these changes affect air quality and the production of allergens, contributing to an increase in asthmatic episodes and allergic respiratory diseases [9,10].

In Brazil, asthma represents a significant public health problem due to limitations in access to standardized treatment, which overloads the capacity for hospital admissions, as evidenced in several studies [11–13]. Several studies studying the correlation between environmental variables and respiratory diseases have been carried out in different cities in Brazil [14–17]. An epidemiological analysis carried out between 2016 and 2020 revealed a significant reduction in the number of hospitalizations for asthma during the COVID-19 pandemic (2019–2020), although the northeast region of Brazil recorded the highest number of hospitalizations in the 2016–2020 period, with 2248 deaths associated with the disease during the period [18]. According to these authors, the higher number of asthma hospitalizations may be associated with increased temperatures and lower latitudes.

Furthermore, when all mentions of asthma on death certificates are considered, in 2000, there was an increase of approximately 50% in mortality rate, ranging from 36% in the state of Ceará to 76% in the state of Rio Grande do Sul, with women and the elderly showing the highest rates [19]. The prevalence of asthmatic symptoms also stands out, with significant variations observed between Brazilian states; for example, the prevalence of active asthma was 14.8% in Maceió, a city on the coastal strip of the northeast region with a hot and rainy climate, compared to 30.5% in Vitória da Conquista, a city in the interior of Northeast Brazil with a high-altitude humid climate [20]. These data indicate not only the severity of the condition, but also the urgent need for monitoring and effective

interventions, especially in regions such as the northeast, where asthma research is still scarce and needs greater attention.

The use of machine learning (ML) techniques in asthma research has shown promise, allowing for more in-depth analysis of complex and varied data. ML involves the use of algorithms to process large volumes of data, identifying patterns without the need for explicit programming by humans [2]. Khanam et al. [1] conducted a scoping review on the application of ML in asthma-related research, identifying 102 relevant articles. The main areas of focus included the prediction of asthmatic episodes (24.5%), asthma phenotype classification (16.7%), and asthma genetic profiling (12.7%). Most of the studies used cohort designs (52.9%) and techniques such as neural networks, clustering, and random forests, which were applied in 20.6%, 18.6%, and 17.6% of the studies, respectively. This approach stands out compared to conventional statistical methods, which often rely on human experience to discover causal relationships and can be subject to error [21].

Despite the increase in the use of ML, few studies have considered variables such as the residential location of individuals [1], and the application of ML techniques to subsidize public policies in the health sector is still scarce; thus, resources for this purpose could be managed more effectively. The northeast region of Brazil tends to have the highest number of registered asthma cases [11]. In this sense, the present study analyzes more than twenty years of data on asthma notifications recorded in the city of Maceió, the capital of the state of Alagoas, located on the east coast of northeastern Brazil, a municipality with approximately 1 million inhabitants and one of the lowest human development indices among Brazilian state capitals, at 0.721 [22]. Five different ML techniques were used to assess the predictability of monthly asthma hospitalizations in order to identify the most effective method for building a hospitalization prediction model based on the models' performance in identifying the relationship between the meteorological and air pollution variables most relevant to asthma records.

## 2. Materials and Methods

### 2.1. Asthma Data

Data on hospitalizations for asthma in Maceió, Alagoas, were obtained from the SUS Statistics Department [23]. The SUS Hospital Information System (SIH/SUS), managed by the Ministry of Health in conjunction with the state and municipal health departments, sends the information from the Hospital Admissions Authorization (AIH) to Datasus, which analyses, processes, and makes these data available at <https://datasus.saude.gov.br/> (accessed on 10 September 2024). The tabulation is based on information related to the pathology according to the International Classification of Diseases (ICD-10), specifically Chapter X, which deals with diseases of the respiratory system, focusing on the morbidity categories associated with asthma, identified by the codes J45 and J46. This study did not require submission to the Research Ethics Committee, as the data are in the public domain and accessible on the internet. In this study, we used monthly time series of asthma hospitalizations from 1998 to 2023.

### 2.2. Pollution Data

The Copernicus Atmosphere Monitoring Service (CAMS) offers a global near-real-time (NRT) service that provides daily analysis and forecasts of reactive trace gases, greenhouse gases, and aerosol concentrations. This modeling system incorporates meteorological and atmospheric composition observations, which are assimilated by ECMWF's 4D-Var assimilation model to create a reanalysis of atmospheric composition. Three recent reanalyses have been published: the Monitoring Atmospheric Composition and Climate (MACC) reanalysis for the years 2003–2012 [24], the “CAMS interim Reanalysis” (CAMS-iRean) for

the years 2003–2018 [25], and the “CAMS Reanalysis” (CAMS-Rean) for the years 2003 to the present [26]. Detailed information on the CAMS reanalysis can be found in [27]. We used the following variables: ozone ( $\text{O}_3$ — $\text{kg} \times \text{kg}^{-1}$ ), sulfur dioxide ( $\text{SO}_2$ — $\text{kg} \times \text{kg}^{-1}$ ), carbon monoxide ( $\text{CO}$ — $\text{kg} \times \text{kg}^{-1}$ ), nitrogen dioxide ( $\text{NO}_2$ — $\text{kg} \times \text{kg}^{-1}$ ) at  $0.75^\circ \times 0.75^\circ$  spatial resolution, and monthly temporal resolution at surface level from 2003 to 2023.

### 2.3. Meteorological Data

The meteorological data used in this study were obtained from the National Meteorological Institute’s conventional weather station located in the city of Maceió (lat:  $-9.55^\circ$ ; lon:  $-35.77^\circ$ ; station code 82994), covering the period from 1998 to 2023 with monthly resolution. The variables recorded included evaporation (mm), representing the amount of water evaporated; sunshine (hours), corresponding to the total hours of sunshine in the month; cloudiness (tenths), reflecting the average level of cloud cover; total monthly precipitation (mm); atmospheric pressure (hPa); wind speed (m/s); relative humidity (%); and maximum, minimum, and average monthly temperatures ( $^\circ\text{C}$ ). Two indirect variables were obtained from the aforementioned data and also made up the list of input data for the ML models: the number of days with rainfall in the month, indicating the frequency of days with rainfall, and potential evapotranspiration (mm), calculated using the Penman–Monteith method.

### 2.4. Machine Learning Models

#### 2.4.1. Random Forest

Random forest (RF) is a machine learning method that operates by building a multitude of decision trees at the time of training and producing the class that is the mode of the classes (classification) or the average of the predictions (regression) of the individual trees [28,29]. RF is a type of ensemble learning, which is a way of combining the predictions of multiple ML models to produce a more accurate and robust result than the individual models. This method is an adaptation of decision trees, where the model generates predictions from a sequence of basic models, as described in Equation (1) [30], where each basic model is a decision tree and  $k$  denotes the number of decision trees [31]:

$$g(x) = f1(x) + f2(x) + f3(x) + \dots + fk(x) \quad (1)$$

Random forests create a set of decision trees, each trained on a random subset of the training data. This process is known as bagging or bootstrap aggregation [32]. Each tree is trained on a random sample of the data, obtained with replacement, called a bootstrap sample. In addition, when building trees, the best split for each node is chosen from a random subset of features, rather than all features [33]. This approach ensures that the trees are diverse and can capture a wide range of patterns in the data. Since each tree is built using different subsets of the data and features, the random forest as a whole is less prone to overfitting than a single decision tree. The final result is obtained by aggregating the predictions from each tree, which reduces variance and produces a more stable model.

RF can cope well with input data that have missing values, and is generally robust to outliers. It can be applied to large, high-dimensional datasets. An additional advantage is that the method can provide estimates of the importance of variables. This is undertaken by looking at how much the prediction error increases when the data for a variable are randomized. Therefore, RF is considered one of the most versatile and easy-to-use algorithms, offering good classification or regression performance without the need for the extensive fine-tuning of hyperparameters, which makes it a reliable starting model for many ML problems.

#### 2.4.2. XGBoost

eXtreme Gradient Boosting (XGBoost) is an advanced and efficient software library for predictive modeling. XGBoost is a tree ensemble technique that uses a gradient descent approach to improve weak learning models by adding a regularization term in the loss function to smooth the adjusted weights and prevent overfitting [34]. XGBoost can be calculated as follows (Equation (2)):

$$\hat{y} = \frac{1}{K} \sum_{k=1}^K f_k(x), f_k \in \Gamma \quad (2)$$

where  $f_k$  represents the output of the  $k$ -th tree,  $x$  is the input vector,  $\Gamma$  denotes the function space containing all possible regression trees, and  $\hat{y}$  is the projected output [35,36].

XGBoost has stood out in many data science competitions for its performance and accuracy capabilities. XGBoost uses gradient boosting, which is an ensemble technique. Instead of building a single model, gradient boosting builds many models in sequence, where each new model tries to correct the errors of the previous model. XGBoost includes regularization terms (L1 and L2) in the calculation of the loss function, which helps to avoid overfitting, i.e., when a model overfits the training data and performs poorly on new data. It can automatically learn the best way to handle missing values during training, as well as the ability to stop training the model as soon as the performance on the validation data starts to deteriorate, thus avoiding wasting computational resources.

#### 2.4.3. Multiple Linear Regression

Multiple Linear Regression (MLR) is one of the most widely used statistical and machine learning approaches to explore relationships between variables, allowing the modeling of linear relationships between a dependent variable and two or more independent variables. According to [37], MLR is widely used to model linear relationships between a dependent variable and several independent variables. According to the authors, the appropriate choice of variables and the regression model is essential to ensure valid statistical results in observational studies, since incorrect selection can lead to inaccuracies in the results. In the reviewed studies, the accuracy and predictive ability of the models were assessed through comparisons between predicted and observed values, highlighting the importance of the careful choice of the type of regression to ensure reliable predictions [37].

#### 2.4.4. Support Vector Regression Model

In 1995, the support vector machine (SVM) algorithm was introduced as a highly efficient method for pattern recognition [38]. The SVM is a supervised learning algorithm used for classification and regression tasks, and has been gaining prominence for its effectiveness in several engineering applications [39]. In this method, data are represented as points in an  $n$ -dimensional space, where the number of dimensions corresponds to the number of features of the samples. The main objective is to identify the ideal hyperplane that separates the different classes of data, minimizing the error. Support Vector Regression (SVR) is a specific variant of SVM, aimed at regression problems, and is able to deal with continuous values. Since its creation, SVR has been widely adopted in several systems and applications [33,40,41].

#### 2.4.5. K-Nearest Neighbors

The  $k$ -nearest neighbors (KNN) algorithm, also known as the memory-based classification method, is a simple and intuitive technique in data mining [42]. In it, the difference between features is calculated using Euclidean distances, enabling work with both continuous



and discrete variables. For example, if the first sample is represented by  $(a_1, a_2, a_3, \dots, a_n)$  and the second by  $(b_1, b_2, b_3, \dots, b_n)$ , the distance  $d$  between them is given by Equation (3):

$$d = \sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2 + \dots + (a_n - b_n)^2} \quad (3)$$

One challenge with using Euclidean distance is that large values can overshadow smaller ones, compromising the analysis. Although KNN works primarily with continuous data, it can also be applied to discrete data. In a hypothetical example, if discrete values in  $a$  and  $b$  differ, this difference is counted as one; if they are equal, the difference is zero [43]. The algorithm is tuned on several neighborhood values until it reaches optimal levels of sensitivity and specificity, as described in the results.

## 2.5. Evaluation Methods for ML Models

In the context of this work, which evaluates the performance of ML models applied to regression problems, the selection of appropriate evaluation metrics is essential for the correct interpretation and validation of predictive models. Metrics such as ROC Curve, F1 Score and Brier Score Loss are widely used in classification problems, but are not appropriate for evaluating regression models. To evaluate such models, specific metrics such as mean absolute error (MAE, Equation (4)), mean squared error (MSE, Equation (5)), root mean squared error (RMSE, Equation (6)), and coefficient of determination ( $R^2$ , Equation (7)) are more appropriate. In the analysis of R-squared, caution is necessary, as it assumes that all variables in the model are independent and may fail to detect non-linear relationships. Moreover, it can produce misleading results in smaller datasets. These metrics quantify the difference between predicted and actual values, providing a direct measure of prediction accuracy in a continuous value context. The same methodological approach has been employed in several studies, such as those by [44–48], and others. The metrics are defined as follows:

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_{i,real} - y_{i,predicted}| \quad (4)$$

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_{i,real} - y_{i,predicted})^2 \quad (5)$$

$$RMSE = \sqrt{MSE} \quad (6)$$

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_{i,real} - y_{i,predicted})^2}{\sum_{i=1}^n (y_{i,real} - \bar{y}_{real})^2} \quad (7)$$

where  $y_{i,real}$  represents the real values and  $y_{i,predicted}$  the predicted values of hospitalizations for asthma, and  $n$  indicates the size of the test dataset.

The coefficient of determination ( $R^2$ ) measures the proportion of the variability of the dependent variable that is explained by the independent variables. It ranges from  $-\infty$  to 1, indicating that the closer to 1, the better the predictive ability of the model, according to the mutual relationship between the ground truth and the prediction model [49]. A negative  $R^2$  means that the predictive performance is lower than the mean model, which can occur when using out-of-sample data or in regressions without intercept.

## 2.6. Autocorrelation Function

Autocorrelation function (ACF) analysis is a statistical technique used to assess the temporal dependence between consecutive observations in a time series [50,51]. In a predictive modeling context, the ACF is particularly useful for examining the structure of model residuals, making it possible to identify possible correlation patterns that may

indicate model inadequacy or the need to include lags in the variables. In the context of the predictive modeling of time series, the ACF plays a crucial role, since the validation of the estimated model depends on the study of its residuals, which is achieved thanks to its ACF [52]. In an adequate model, it is expected that the residuals do not show significant autocorrelation, i.e., that the ACF of the residuals shows values close to zero for all lags within the confidence intervals, indicating independence between the observations over time.

To calculate the ACF, each value in the time series is correlated with the values at different lags, resulting in a series of autocorrelation coefficients. These coefficients are interpreted according to significance levels, usually represented by dashed lines on the ACF graph. Autocorrelation values that exceed these limits indicate the presence of significant temporal dependence for that specific lag. In the case of forecasting model residuals, this dependence may suggest that the model has not adequately captured the entire temporal structure of the data, which may jeopardize the accuracy and reliability of the forecasts.

### 3. Results

#### 3.1. Machine Learning Performance (More Data or More Variables?)

Predicting asthma hospitalizations using ML models is sensitive to both the number of predictor variables and the amount of data available. Firstly, we carried out two experiments to assess the impact of these variables. Experiment 1 (hereafter exp1) includes the additional variables related to air pollution described in Section 2.2, but with a smaller amount of temporal data (January 2003 to December 2023, 252 months). Experiment 2 (hereafter exp2) excludes the pollution variables, but has a longer time series of meteorological variables (January 1998 to December 2023, 312 months). The aim of this comparison was to identify whether more variables with less data or fewer variables with more data result in superior performance.

To ensure the robustness and reliability of the asthma hospital admissions forecasts, standard procedures were followed for the application of each ML model. Firstly, each model was fitted and trained using 80% of the available data, while the remaining 20% was reserved for validation, following standard practice to avoid overfitting and allow an accurate assessment of performance. However, each model was set up with a different 80–20 split, allowing for a more flexible assessment of performance and observation of how each behaves with different training and test samples. This approach helps to minimize the bias associated with a single split, providing a more reliable measure of the models' ability to generalize. Therefore, the figures presenting the observed and predicted data throughout this topic will not have the same time-series length, since each model was fitted with a different 80–20 split.

The models were configured and optimized using the best-practice recommendations for each technique: For the MLR model, tests were applied to verify assumptions of the linearity and normality of the residuals. We did not remove the predictor variables that showed multicollinearity between them, as the main objective of the model is accurate prediction and not interpretation of the coefficients; collinearity is not necessarily a critical problem. Furthermore, it was also observed that removing the collinear variables led to significantly worse performance. In the RF model, the *mtry* hyperparameter (number of variables used at each node) was adjusted to minimize the out-of-bag (OOB) error, while the number of trees was set to balance between accuracy and computational efficiency. In XGBoost, a hyperparameter optimization was performed, adjusting the number of trees, the learning rate, and the maximum depth of the trees, in order to minimize the validation of MSE. For the KNN model, the value of *k* was selected through cross-validation to ensure the best fit to the dataset. In the SVR model, the regularization parameters and the appropriate

kernel were adjusted to balance the complexity of the model and the prediction error. In addition, data normalization and standardization techniques were employed to ensure that the variables were comparable in scale and to avoid biases that could compromise the performance of the models.

The results are shown in Table 1, where the five models used are compared in terms of the *MSE*, *RMSE*, *MAE*, and  $R^2$  metrics. Figure 1 shows the respective scatter plots between the data observed and predicted by the ML models. In addition, the correlations between the values predicted by the models and the observed values are shown in Figure 2, which compares the predictions of both experiments for each model. For the MLR model, exp1 performed better than exp2, with all metrics being superior, indicating that the inclusion of pollution variables improves the model's accuracy. This indicates that the model benefits from the inclusion of air pollution variables even with the loss of part of the time series. The RF model, on the other hand, performed better in exp1, with an *MSE* of 179.07, *RMSE* of 13.38, and *MAE* of 10.62, although exp2 showed a slight increase in  $R^2$  of 0.58. These results indicate that RF works better with a shorter period and with pollution variables, while increasing the period of temporal data did not bring great advantages in terms of accuracy.

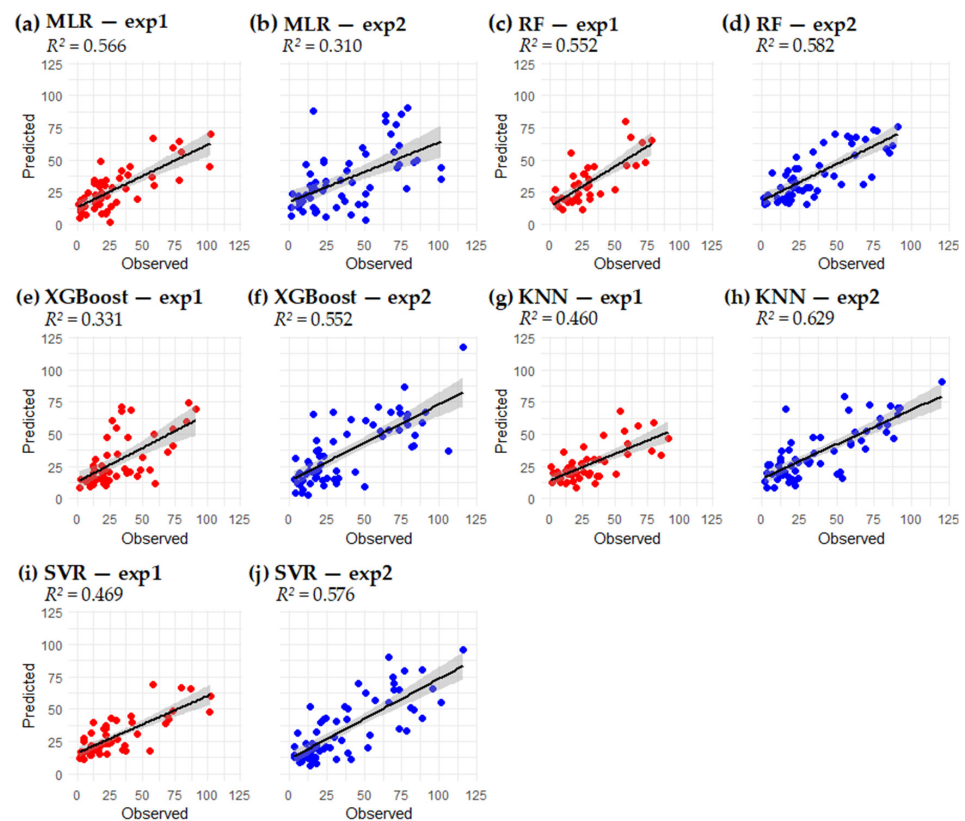
**Table 1.** Statistical metrics for the MLR, RF, XGBoost, KNN, and SVR models in experiment 1 (exp1) and experiment 2 (exp2).

Model	Experiment	<i>MSE</i>	<i>RMSE</i>	<i>MAE</i>	$R^2$
MLR	exp1	298.21	17.27	13.06	0.566
	exp2	582.29	24.13	18.54	0.310
RF	exp1	179.07	13.38	10.62	0.552
	exp2	267.70	16.36	13.28	0.582
XGBoost	exp1	341.61	18.48	14.34	0.331
	exp2	402.35	20.06	15.14	0.552
KNN	exp1	291.30	17.07	12.87	0.460
	exp2	362.01	19.03	15.44	0.629
SVR	exp1	295.88	18.92	13.17	0.557
	exp2	339.66	19.25	13.80	0.611

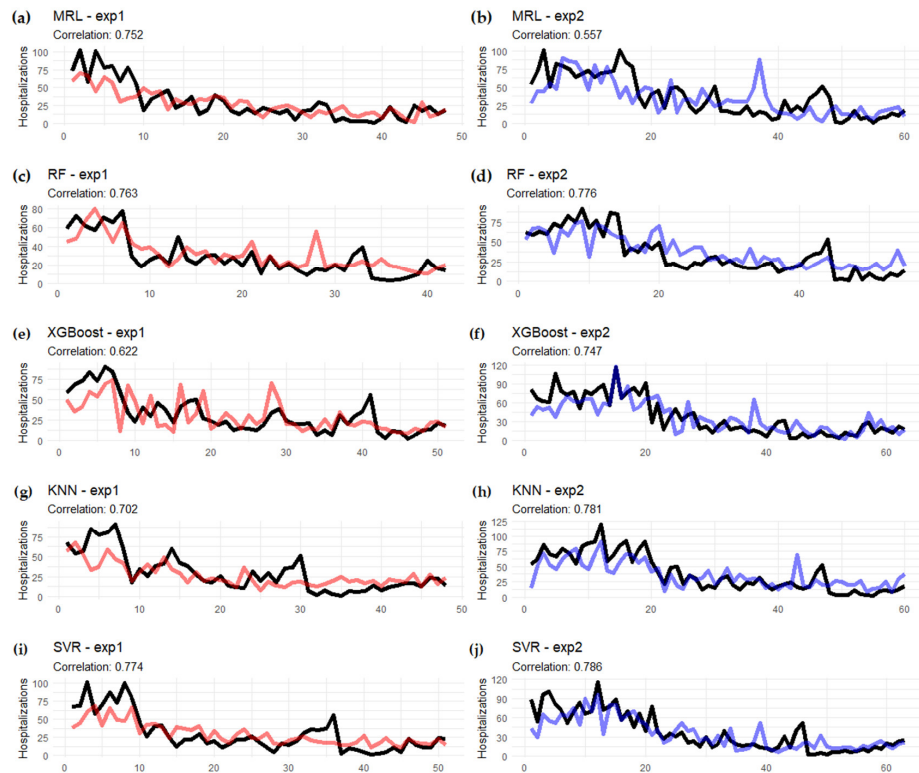
In exp2, the RF model still performed relatively well, especially considering the slightly higher  $R^2$  than in exp1. Thus, the RF model still manages to capture a good amount of the data variability, although with a lower overall accuracy in terms of mean error. XGBoost performed the worst overall among the analyzed models, with relatively high errors in both experiments. In exp1, XGBoost obtained an *MSE* of 341.61, *RMSE* of 18.48, *MAE* of 14.34, and  $R^2$  of 0.331, indicating a low explanatory power and high margin of error in the predictions. In exp2, XGBoost's performance was even worse. These results suggest that XGBoost did not benefit from either the inclusion of air pollution variables or the increase in the time period of data. The KNN model performed averagely, with a superior performance in exp2.

However, considering the *MSE*, *RMSE*, and *MAE* values, KNN was not as effective as the other models tested. Although it was not the model with the lowest errors, KNN stood out in exp2 in terms of  $R^2$ , showing that, despite having higher errors, it is able to explain a significant part of the variability of asthma cases when working with a longer time series. This indicates that, although KNN presents good adjustment capacity in long-term data, it is still surpassed in accuracy by the other models. The SVR model presented a similar performance to KNN, with intermediate error metrics. exp1 recorded an *MSE* of 295.88 and  $R^2$  of 0.56, while exp2 had an *MSE* of 339.66 and  $R^2$  of 0.61. These results indicate that SVR was not as efficient as the other models, but still presented a reasonable adjustment capacity, with a slightly better performance in exp2 in terms of explaining the variability.





**Figure 1.** Scatter plots of observed versus predicted data from MLR (a,b), RF (c,d), XGBoost (e,f), KNN (g,h), and SVR (i,j) models. Data from exp1 in red and exp2 in blue.



**Figure 2.** Observed versus predicted time series of MLR (a,b), RD (c,d), XGBoost (e,f), KNN (g,h), and SVR (i,j) models. Observed data in black, exp1 in red and exp2 in blue.

Considering the error metrics, the model that presented the best performance was RF in exp1. This model obtained the lowest values of  $MSE$  (179.07),  $RMSE$  (13.38), and  $MAE$

(10.62) among all the models and experiments tested. These results indicate that RF was more accurate in predicting asthma cases, since it presented the lowest *MSE*, *RMSE*, and *MAE*, reflecting greater accuracy in predicting values close to those observed. In addition, the  $R^2$  of RF in exp1 was 0.552, which, although not the highest among the models, still indicates a reasonable ability to explain the variance in the data. In the context of this analysis, the lowest mean error is more relevant to identify the model with the best fit, and RF in exp1 clearly stands out in this aspect, being, therefore, the most suitable model among those tested for predicting asthma cases.

The scatterplot presented in Figure 1 compares the observed values of hospitalizations due to asthma with the predicted values for each of the models used. Graphs (a) and (b) correspond to the MLR model, (c) and (d) to RF, (e) and (f) to XGBoost, (g) and (h) to KNN, and (i) and (j) to SVR, with the exp1 dataset (red) represented by the graphs on the left, which include air pollution variables, and the exp2 dataset (blue) by the graphs on the right, which are based exclusively on meteorological data but cover a longer period of time, as previously described. For the MLR (Figure 1a,b), the scatter of the points around the identity line is considerable, especially in exp2, indicating that the MLR has difficulty in capturing the variation in the data accurately, especially for higher values of hospitalizations. In the RF model plots (Figure 1c,d), the points are closer to the regression line in both experiments, and the distribution along the line is more balanced. This indicates that RF captures the fluctuations in the observed data more robustly, especially in mean hospitalization values. In addition, the dispersion is more uniform, which reflects the greater stability of the model in the different value ranges. For XGBoost (Figure 1e,f), a greater dispersion of the points is observed, especially in exp1. This model tends to have a less accurate prediction for extreme values, indicating that XGBoost may be biased by overfitting to specific patterns in the data without generalizing as well as other models. The KNN model, represented in Figure 1g,h, shows a dispersion pattern that approaches the regression line, especially in experiment 2. KNN appears to efficiently capture the relationship between the variables for the second dataset, maintaining a distribution relatively close to the regression line. However, in exp1, some more significant deviations are observed for higher values of hospitalizations, suggesting that KNN may be less robust in situations with less data or greater variability. The SVR model (Figure 1i,j) presents a distribution of points reasonably close to the regression line in both experiments, although with some dispersions standing out at extreme values. The results, in general, show that SVR manages to maintain a good correspondence with the observed values, but also faces difficulty in accurately predicting higher values of hospitalizations, where the dispersion increases. RF and KNN seem to deal better with data variability in the second experiment, while MLR demonstrates more evident limitations in both experiments, particularly for extreme values. XGBoost and SVR, despite presenting good proximity to mean values, show high dispersion for higher values, indicating that they could benefit from additional adjustments.

Figure 2 shows the comparison between the observed and predicted time series for the different models in two experiments: exp1 (left column in red) and exp2 (right column in blue). As explained previously, each machine learning model was fitted with a different 80–20 split of the data, which resulted in variations in the training and validation sets for each one. For this reason, the figures showing the observed and predicted data time series show X and Y axes with different lengths. It is possible to observe the actual hospitalizations (black line) and the model predictions (colored line) for the analysis period of each model, with the correlation between the observed and predicted values indicated in each graph. In MLR, it is observed that the performance in exp1 (Figure 2a) is superior to that of exp2 (Figure 2b) (0.752 versus 0.557). This result indicates that the MLR model fits better when there is a smaller set of variables including air pollution. The prediction in exp2 presents

larger oscillations and a less accurate fit to the fluctuations observed in the hospitalization predictions. The RF model (Figure 2b,c) presents high correlations in both experiments (0.763 in exp1 and 0.776 in exp2). Although both experiments demonstrate good adherence to the behavior of the observed time series, it is noted that the RF model in exp2 presents a slightly better and more consistent fit, suggesting that the RF benefits from the longer time series of exp2. The XGBoost model (Figure 2e,f) presents moderate performance in exp1 (correlation 0.622), improving significantly in exp2 (correlation 0.747). This result indicates that XGBoost also benefits from a longer time series. However, there are still deviations in some peaks of hospitalizations, suggesting that XGBoost has some difficulty in capturing some more abrupt fluctuations. KNN (Figure 2g,h) shows an improvement in correlation from 0.702 in exp1 (Figure 2g) to 0.781 in exp2 (Figure 2h). This result suggests that KNN is favored by the larger amount of temporal data in experiment 2, allowing a more robust fit to hospitalization patterns. SVR presents a high correlation in both exp1 (0.774 in Figure 2i) and exp2 (0.786 in Figure 2j), with a slight increase in the second experiment. SVR is the model that presents greater stability between the two experiments, with predictions that closely follow the observed series. The high correlation in both experiments suggests that SVR is effective in capturing time series trends, regardless of the number of variables or the length of the time series. Overall, the time series plots in Figure 2 highlight that the RF, KNN, and SVR models show greater fitting ability in exp2, benefiting from the longer time series and better capturing hospitalization patterns over time. Among the models, SVR stands out for its consistency across both experiments, while XGBoost and MLR show more limited performance, especially in exp1. These observations indicate that for time series with hospitalization data and atmospheric variables, models such as RF, KNN, and SVR may be better suited to capture temporal variations when a longer time series is available.

### 3.2. Importance of Variables in Tree Models

As the RF model proved to be the best of the five ML models analyzed, we will now evaluate the importance of the variables according to the trained RF model. The importance of the variables is measured by the increase in the purity of the nodes that the variable produces, which is a sum of the decrease in the Gini index (a measure of impurity) that each variable brings to the splits it creates in all the decision trees in the forest, so it is useful for assessing the role of the variables in the decision structure and in splitting the data within the model. We also evaluated the “% Increase in Mean Squared Error” (%IncMSE), a metric that reflects the percentage increase in MSE that occurs when the values of a specific variable are randomly permuted while all other variables remain unchanged. In other words, a higher %IncMSE indicates that the variable is more important for the model’s prediction, because when its information is corrupted, the model’s error increases significantly.

Table 2 shows the numbers that are the sum total of the decrease in impurity that this variable provided in the trees of the random forest in experiments 1 and 2. For the RF model in exp1, the variables with the greatest contribution to the model’s accuracy, as measured by the percentage increase in %IncMSE, were minimum temperature (36.03%), SO<sub>2</sub> (20.82%), and insolation (14.60%). These variables showed great importance in the forecast, suggesting that minimum temperature and SO<sub>2</sub> are strong indicators of the response variable in the context of climate and pollution data.

**Table 2.** Gini index and the percentage increase in MSE for each variable in the random forest model in experiments 1 and 2.

Variables	% Increase in Mean Squared Error		Increase in Node Purity	
	exp1	exp2	exp1	exp2
Evaporation	10.10	11.10	6341.92	15,129.60
Evapotranspiration	4.67	11.85	2069.43	10,105.40
Insolation	14.60	8.01	9912.74	11,661.47
Cloudiness	2.92	12.59	2909.33	12,526.92
Days with precipitation	4.24	8.41	1808.69	6677.03
Precipitation	2.29	5.68	2367.85	8299.80
Atmospheric pressure	8.75	20.68	6430.07	27,068.95
Maximum temperature	2.53	8.90	1708.85	8371.45
Average temperature	4.75	14.94	4537.50	16,793.87
Minimum temperature	36.03	46.17	36,416.31	78,485.94
Relative humidity	7.10	11.72	3712.47	12,971.09
Wind speed	0.92	6.02	1404.11	7181.92
CO (kg kg <sup>-1</sup> )	4.63	-	3788.71	-
NO <sub>2</sub> (kg kg <sup>-1</sup> )	7.35	-	4386.75	-
O <sub>3</sub> (kg kg <sup>-1</sup> )	1.10	-	2647.99	-
SO <sub>2</sub> (kg kg <sup>-1</sup> )	20.82	-	13,113.98	-
PM10 (kg m <sup>-3</sup> )	3.83	-	3269.54	-
PM2.5 (kg m <sup>-3</sup> )	3.43	-	2511.66	-

As for the node purity metric, which measures the reduction in impurity caused by splitting a node using a specific variable, again the minimum temperature (36,416.31) was the most important variable, followed by SO<sub>2</sub> (13,113.98) and insolation (9912.74), reinforcing its substantial impact on the accuracy of the predictions.

In exp2, which incorporates a longer time series without the pollution variables, it was observed that minimum temperature again stood out as the most important variable with an %IncMSE of 46.17%, followed by atmospheric pressure (20.68%) and average temperature (14.94%). However, the importance of node purity increased even more for minimum temperature (78,485.94), while pressure (27,068.95) and average temperature (16,793.87) also showed high relevance. These values indicate that, in the absence of the air pollution variables, the importance of variables such as atmospheric pressure and average temperature increased significantly, reflecting changes in the relative importance of the predictor variables.

In terms of interpretation, the differences between experiments 1 and 2 suggest that the inclusion of a longer time series may alter the importance of certain variables in the RF model. While the minimum temperature variable remains the most relevant variable in both experiments, variables such as atmospheric pressure and average temperature become more relevant in the RF model in exp2, suggesting that these variables may capture trends and fluctuations that are more evident on an extended time scale. It is clear that the minimum temperature variable seems to have the greatest predictive power for the number of asthma cases. This does not necessarily mean that the minimum temperature variable causes more asthma cases, but rather that there is a strong association between it and the number of cases within the dataset supplied to the models. A strong association does not imply causality and these influences should be analyzed in more detail with further studies to understand the cause-and-effect relationship.

### 3.3. Random Forest Prediction

We evaluated different versions of the RF model with and without lags, and incorporated the variable “asthma cases” as a predictor variable with and without time lags. The justification for this is due to the importance of capturing the temporal effects and inertia of the time series, especially in a context where past variables can directly influence future forecasts.

The first RF model (hereafter referred to as RF1) only considers the climate and pollution variables observed at the current time (lag 0) to predict hospitalizations. By not using lags, it makes the forecast based on the contemporary values of the climate and pollution variables. This is useful for understanding the model's ability to predict the number of asthma cases based on current factors, ignoring any temporal dependence that may exist.

The second RF model (hereafter referred to as RF2) uses climate and pollution variables with lag 1, excluding lag 0 variables, and does not use the variable number of hospitalizations with lag 1 as a predictor variable. In this model, the aim is to understand whether the climate and pollution conditions of the previous month have a direct impact on asthma cases in the current month. This type of analysis is especially relevant in public health, where previous environmental factors can have a delayed effect on health events.

The third RF model (hereafter referred to as RF3) explores climate and pollution variables with lag 1, excluding lag 0 variables, and uses the variable number of hospitalizations with lag 1 as a predictor variable. This model is similar to RF2, but by adding the number of hospitalizations with lag 1 as a predictor variable, it allows the model to capture the temporal autocorrelation in the asthma cases themselves, i.e., the influence of past cases on future cases. In time-series models, it is common for the target variable to have a correlation with its previous values, especially in seasonal phenomena or those with temporal inertia.

The fourth model is the RF combining climate and pollution variables with lag 0 and lag 1, as well as including the variable number of hospitalizations with lag 1 as a predictor variable. This model makes the most of the potential information, assessing both the immediate and delayed effects of environmental and pollution variables, as well as the impact of past hospitalizations. This model is particularly useful if the combination of current and lagged variables has a synergistic effect, i.e., if the effect of climate and pollution conditions is both immediate and prolonged.

Table 3 shows the values of the performance comparison metrics between the four RF models tested. The results highlight the RF4 model as the best configuration for predicting the number of hospitalizations for asthma, highlighting the importance of including variables with a time lag and the autocorrelation of the hospitalizations variable as a predictor. When comparing the four approaches, we observed that the models using variables with a lag of 1 month, especially with the inclusion of the lagged hospitalizations variable itself, show superior performance.

**Table 3.** Statistical metrics for random forest models 1 (RF1), 2 (RF2) 3 (RF3), and 4 (RF4).

Model	MSE	RMSE	MAE	R <sup>2</sup>
RF1	233.584	15.283	11.854	0.527
RF2	256.768	16.024	12.384	0.480
RF3	142.158	11.923	8.937	0.712
RF4	116.514	10.794	8.255	0.764

Table 4, similar to Table 2, shows the total sum of the reduction in impurity provided by this variable in the four RF models. The results indicate that the inclusion of variables with lag, especially the history of hospitalizations, significantly enriches the predictive capacity of the models. The autocorrelation of hospitalizations is essential for capturing consistent temporal patterns, while the combination of variables with and without lag provides a more comprehensive view of the environmental factors that contribute to fluctuations in asthma hospitalizations. RF4, the best model according to the performance metrics, which incorporates both contemporary and lagged variables, including the number of hospitalizations with lag 1, shows a more balanced scenario in terms of the importance of the variables. The model manages to capture both the immediate and delayed effects



of environmental conditions and hospitalizations. The lag 1 hospitalizations variable remains one of the most important, but climate and pollution variables (such as minimum temperature and SO<sub>2</sub>) also gain relevance when they are available on both time scales. This balance indicates that the combination of present and past information optimizes the model's ability to capture the complex factors that influence hospitalizations.

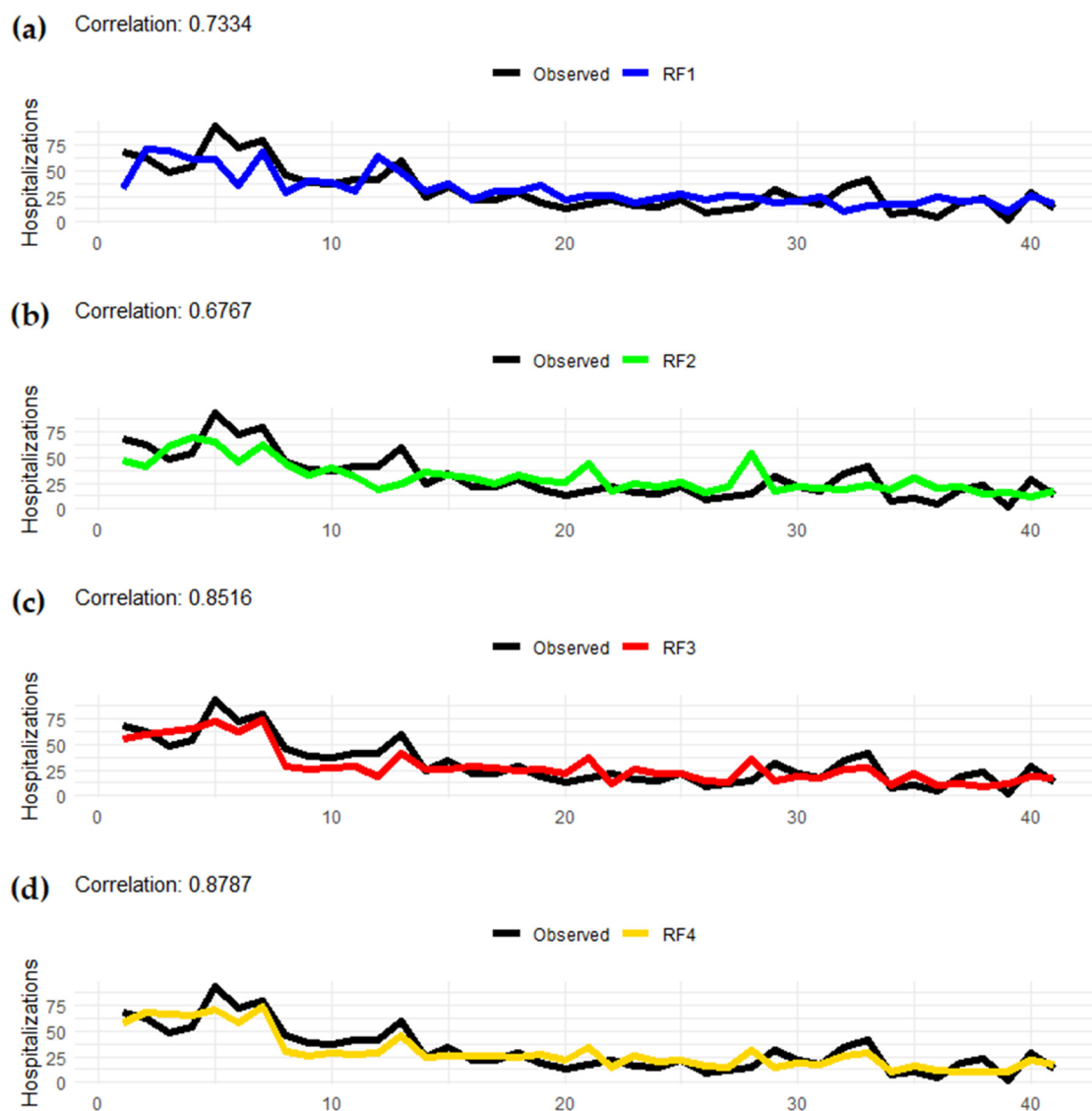
**Table 4.** Gini index and the percentage increase in MSE for each variable for random forest models 1 (RF1), 2 (RF2), 3 (RF3), and 4 (RF4). The acronym %IMSE indicates increase in mean squared error (%IncMSE) and INP indicates node purity (IncNodePurity).

Variables	RF1		RF2		RF3		RF4	
	%IMSE	INP	%IMSE	INP	%IMSE	INP	%IMSE	INP
Evaporation (lag 0)	10.2	6733.2	-	-	-	-	4.0	2380.2
Evapotranspiration (lag 0)	5.2	1849.4	-	-	-	-	2.6	373.2
Insolation (lag 0)	8.9	3369.0	-	-	-	-	4.2	963.7
Cloudiness (lag 0)	3.1	2386.9	-	-	-	-	2.0	524.6
Days with precipitation (lag 0)	2.9	1545.7	-	-	-	-	2.8	417.4
Precipitation (lag 0)	1.7	1476.2	-	-	-	-	0.1	425.1
Atmospheric pressure (lag 0)	8.6	6720.3	-	-	-	-	0.8	1170.6
Maximum temperature (lag 0)	3.9	1881.9	-	-	-	-	2.4	465.5
Average temperature (lag 0)	8.1	5069.5	-	-	-	-	3.5	1067.0
Minimum temperature (lag 0)	16.8	13,037.0	-	-	-	-	9.2	5273.3
Relative humidity (lag 0)	3.4	2511.5	-	-	-	-	1.5	631.9
Wind speed (lag 0)	2.6	1303.1	-	-	-	-	1.2	309.7
CO (lag 0)	3.9	3442.6	-	-	-	-	0.9	814.2
NO <sub>2</sub> (lag 0)	8.4	3286.9	-	-	-	-	1.8	832.0
O <sub>3</sub> (lag 0)	4.8	2409.4	-	-	-	-	3.1	455.2
SO <sub>2</sub> (lag 0)	16.1	8093.4	-	-	-	-	6.9	2580.8
PM10 (lag 0)	3.8	2923.0	-	-	-	-	2.8	805.2
PM2.5 (lag 0)	4.8	2442.4	-	-	-	-	1.8	654.0
CO (lag 1)	-	-	3.1	3347.8	2.2	1600.0	1.4	666.6
NO <sub>2</sub> (lag 1)	-	-	7.2	3170.7	3.3	1604.0	1.4	858.2
O <sub>3</sub> (lag 1)	-	-	2.4	2032.2	3.0	1100.3	1.2	506.9
SO <sub>2</sub> (lag 1)	-	-	16.7	7796.1	7.3	3645.4	5.5	2735.3
PM10 (lag 1)	-	-	2.4	1869.1	2.2	1318.7	2.5	675.9
PM2.5 (lag 1)	-	-	2.4	2010.2	1.8	1560.7	4.0	727.7
Evaporation (lag 1)	-	-	6.7	5170.4	4.9	3388.1	4.0	1500.4
Evapotranspiration (lag 1)	-	-	4.6	2075.2	0.0	1013.1	2.1	385.0
Insolation (lag 1)	-	-	5.9	3060.8	3.6	1617.4	1.8	560.8
Cloudiness (lag 1)	-	-	4.1	3049.7	2.9	1895.9	2.7	715.8
Days with precipitation (lag 1)	-	-	3.2	1943.5	1.5	1305.7	2.9	482.4
Precipitation (lag 1)	-	-	4.5	2352.3	1.8	1280.9	-0.6	677.1
Atmospheric pressure (lag 1)	-	-	10.3	6871.8	5.2	3476.9	2.4	2387.5
Maximum temperature (lag 1)	-	-	0.5	2139.2	4.7	1174.7	2.7	583.5
Average temperature (lag 1)	-	-	8.7	5105.7	6.0	3569.8	4.8	1639.9
Minimum temperature (lag 1)	-	-	18.4	13,908.0	10.5	7307.2	9.1	4912.3
Relative humidity (lag 1)	-	-	3.5	2551.7	1.6	1391.7	1.5	676.5
Wind speed (lag 1)	-	-	0.9	1110.6	1.9	650.7	1.1	250.7
Hospitalizations (lag1)	-	-	-	-	39.9	32,661.0	36.8	31,592.1

In the RF1 model, with a similar configuration to RF-exp1, it continued to show that minimum temperature and SO<sub>2</sub> are very important both in terms of increasing the mean squared error and in terms of node purity. In RF2, using only variables with lag 1, but without including the lag of the number of hospitalizations, there is a general decrease in the importance of the variables compared to RF1, although some variables, such as minimum temperature with lag 1 and SO<sub>2</sub> with lag 1, continue to be relevant. The RF3 model, a configuration similar to RF2, but which adds the lagged hospitalization variable (hospitalizations lag 1) as a predictor, stands out for a substantial improvement in the importance of this variable in the model. This result shows that the use of the lagged hospitalization variable allows the model to better capture temporal dependencies, indicating that past hospitalization values are robust predictors of future values. Finally, RF4, which incorporates both contemporary and lagged variables, including hospitalizations with lag 1, shows a more balanced scenario in terms of the importance of the variables. The model

manages to capture both the immediate and delayed effects of environmental conditions and hospitalizations. The variable including hospitalizations with lag 1 remains one of the most important, but climate and pollution variables (such as minimum temperature and  $\text{SO}_2$ ) also gain relevance when they are available on both time scales.

Figure 3 shows a comparison between the observed values of asthma hospitalizations (black line) and the values predicted by the four RF models: RF1 (without lagged variables), RF2 (with lagged climate and pollution variables, but without the variable hospitalization with lag 1), RF3 (with lagged climate and pollution variables, in addition to the variable hospitalization with lag 1), and RF4 (including both lagged and non-lagged variables, including hospitalization with lag 1). Each figure (Figure 3a–d) shows the correspondence between the models' predictions and the observed time series, and also indicates the correlation between the predicted and observed values.

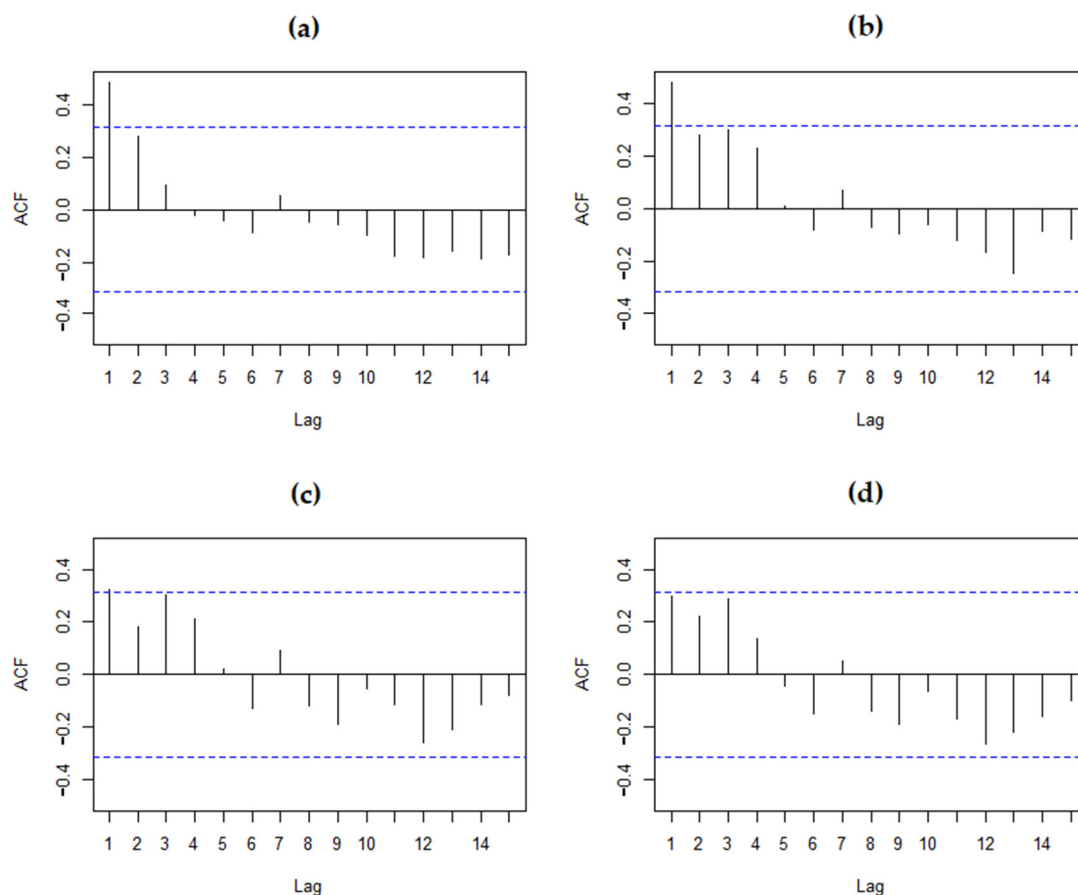


**Figure 3.** Observed (black line) versus predicted time series of the random forest models 1 (a) RF1, blue line, 2 (b) RF2, green line, 3 (c) RF3, red line, and 4 (d) RF4, yellow line.

Corroborating the previous results, the RF4 model (Figure 3d), which combines variables with and without lag, including the variable hospitalizations with lag 1, shows the highest correlation of all the models, with a value of 0.8981. This model is able to capture both the immediate and delayed effects of environmental variables and historical hospital-

ization data. The inclusion of variables on both time scales allows the RF4 model to better adjust the conditions that influence hospitalizations, providing predictions that are closer to the observed values. However, although the RF3 model (Figure 3d) is not superior to RF4, it performs substantially well, with a correlation of 0.8723 between the predicted and observed values. The use of the variable hospitalizations with lag 1 considerably improves the model's fit, capturing the peaks and valleys of the time series more accurately. This result highlights the importance of autocorrelation in hospitalization data, indicating that the history of hospitalizations is a robust predictor of future hospitalization behavior.

The graphs in Figure 4 show the autocorrelation function (ACF) of the residuals for the four RF models (RF1, RF2, RF3, and RF4). The autocorrelation function is a measure of how much a data series is correlated with itself at different lags. In general, the absence of significant autocorrelation in the residuals is an indication that the model is capturing the temporal dependencies of the data well, while the presence of autocorrelation suggests that important aspects of the temporal pattern may be being overlooked.



**Figure 4.** Autocorrelation of the residuals from random forest models 1 (RF1—(a)), 2 (RF2—(b)), 3 (RF3—(c)), and 4 (RF4—(d)).

Due to the use of variables with lag 1, and especially the variable of hospitalizations with lag 1 itself as a predictor in the models, it became necessary to carefully evaluate the autocorrelation functions of the residuals. The inclusion of a variable of interest (such as hospitalizations) with lag allows the model to capture possible autocorrelated temporal dependencies, in which the number of hospitalizations in one period is directly related to the number of hospitalizations in previous periods. This is common in phenomena where past behavior influences future behavior, such as in public health time series and environmental data. However, the use of lagged variables can introduce spurious correlations, since there is a possibility that the model learns patterns in the historical data but

fails to capture new temporal variations not explained by the lag of the variables. This implies that if autocorrelation is still present in the model residuals, it means that there are temporal patterns that have not been adequately captured, suggesting a possible need for additional model adjustments. Thus, evaluating the autocorrelation functions of the residuals makes it possible to check whether the model is in fact capturing all the relevant temporal dependencies. The absence of autocorrelation in the residuals is an indication that the model is well adjusted and that the variables with lag 1 and the hospitalizations variable itself are fulfilling their predictive role, leaving no unexplained temporal patterns. Therefore, the analysis of residual autocorrelation is essential to ensure that the model is well specified and to avoid misinterpretation of the results, especially in time series where historical behavior plays an important role in future forecasting.

The results of the autocorrelation functions of the residuals (Figure 4) indicate that models RF3 and RF4 perform best with non-autocorrelated residuals, suggesting that they capture the temporal structure of the data well. In addition to the lagged variables, the number of previous hospitalizations was crucial to capturing temporal dependencies and improving the quality of the predictions, since the inclusion of the lagged target variable allowed the model to adjust not only to climate and pollution factors, but also to the temporal inertia in asthma cases, which is typical of phenomena with dependence on their past values. Thus, for future studies, modeling time series with climate and health variables with lag is recommended, particularly when using data with strong temporal dependence, as is the case with asthma hospitalizations. The fact that the RF1 and RF2 models only show positive autocorrelations in the first lag suggests that they both capture the immediate structure of the data to some extent, but are unable to completely model the temporal dynamics. The presence of autocorrelation in the first lag indicates that there is some temporal dependence in asthma hospitalization cases that occurs from one period to the next. However, as the autocorrelations are not significant in subsequent lags, this suggests that these models are only capturing part of the immediate temporal effect, without predicting longer-term relationships.

#### 4. Summary and Conclusions

In this study, we carried out a comprehensive evaluation of multiple ML models for predicting asthma hospitalizations based on a variety of meteorological and pollution variables. Two experiments were conducted to assess the impact of using datasets with different durations. Exp1 used a shorter time series and a larger set of variables, while Exp2 incorporated an extended time series and fewer variables. This approach allowed us to analyze the predictive power of models trained with datasets of different sizes and temporal coverage, highlighting the balance between data richness and model performance.

The results showed that the model with the best performance was the RF model using a larger set of variables, even with a shorter time series. This result highlights the importance of the diversity and richness of predictor variables over the length of the time series in some cases. The inclusion of a greater number of variables, such as air pollutants in this study, allowed the RF model to capture patterns that the other models found more difficult.

The analysis of the importance of the variables within the RF models revealed interesting insights. Minimum temperature and SO<sub>2</sub> stood out as influential predictors in all the models analyzed. These results are in line with previous studies correlating asthma cases with exposure to different air pollutants. Achakulwisut et al. [53] estimate that 13% of new pediatric asthma cases per year can be attributed to exposure to air pollutants such as NO<sub>2</sub>. Ding et al. [54] showed that NO<sub>2</sub> plays a key role in asthma attacks in Chongqing, China. Razavi-Termeh et al. [55] indicated that different pollutants have more significant effects depending on the season, with PM<sub>2.5</sub> being the most prominent pollutant. Alarming

data indicate that global life expectancy has been reduced by at least one year due to high concentrations of pollutants [56]. The prominence of minimum temperature can be attributed to its role in respiratory health, as low temperatures are known to exacerbate asthma symptoms and increase vulnerability to respiratory problems. SO<sub>2</sub>, a common pollutant from industrial processes, has also well-documented respiratory effects [57–65], potentially triggering asthma attacks and aggravating symptoms. These findings highlight the importance of including environmental conditions, especially those directly related to respiratory health, as key factors in predictive models for asthma hospitalizations.

Testing different RF model configurations provided interesting insights into the importance of temporal dependencies in the data. RF1, which used only non-lagged variables, demonstrated that current environmental conditions have predictive value but lack the ability to fully capture temporal patterns. RF2, which used only lagged environmental variables without the lagged hospitalizations variable, performed slightly worse, suggesting that past environmental conditions do not play a relevant role in predicting current hospitalizations. While relying on lagged variables reflects current data availability constraints, incorporating real-time data in future studies could enhance predictive accuracy and public health applicability. However, RF3 and RF4, which included the lagged hospitalizations variable, showed the best performance, indicating that recent trends in hospitalizations contribute significantly to predicting future cases. The superior performance of RF4 confirms that the combination of lagged and non-lagged data allows the model to take advantage of both immediate and historical factors, improving its predictive capacity. Although RF4 stood out as the best model, RF1 still offers practical value. RF1, being simpler and less computationally demanding, can be considered in situations where a more immediate model is required, based only on current conditions.

This study demonstrates the predictive potential of RF models for predicting asthma hospitalizations based on environmental data. The importance of variables such as minimum temperature and SO<sub>2</sub> highlights the relevance of climate and pollution factors in respiratory health outcomes. Furthermore, our exploration of various RF configurations highlights the need to incorporate temporal dependencies to achieve optimal predictions, with RF showing excellent results in a more comprehensive approach. Future studies could explore further refinements, including other lag structures and ML algorithms, and explore additional environmental variables to continue improving predictive accuracy and our understanding of environmental impacts on asthma hospitalizations.

**Author Contributions:** Conceptualization, J.S.d.R., R.L.C. and F.D.S.S.; methodology, J.S.d.R., R.L.C. and F.D.S.S.; software, J.S.d.R., R.L.C. and F.D.S.S.; validation, R.L.C. and F.D.S.S.; formal analysis, F.D.S.S.; data curation, R.L.C. and F.D.S.S.; writing—original draft preparation, J.S.d.R., R.L.C. and F.D.S.S.; writing—review and editing, J.S.d.R., R.L.C., F.D.S.S., E.D.F.d.S., T.R.C., R.H.C., S.R.M.V., D.J.D.N., J.F.S.F., C.E.C.B., J.B.C.J., H.S.d.R., K.R.M., M.C.C.L., T.R.F., M.H.G.d.S.V., M.F.A., G.L.M., H.B.G. (Heliofábio Barros Gomes) and H.B.G. (Helber Barros Gomes); visualization, J.S.d.R., R.L.C., F.D.S.S., E.D.F.d.S., T.R.C., R.H.C., S.R.M.V., D.J.D.N., J.F.S.F., C.E.C.B., J.B.C.J., H.S.d.R., K.R.M., M.C.C.L., T.R.F., M.H.G.d.S.V., M.F.A., G.L.M., H.B.G. (Heliofábio Barros Gomes) and H.B.G. (Helber Barros Gomes). All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Data Availability Statement:** The original data presented in the study are openly available at <https://bdmep.inmet.gov.br/> (accessed on 10 September 2024) and <https://www.ecmwf.int/en/forecasts/dataset/cams-global-reanalysis> (accessed on 10 September 2024).

**Acknowledgments:** The first and second authors thank the "Conselho Nacional de Desenvolvimento Científico e Tecnológico—(CNPq)" for the financial support during the conception of this study.

**Conflicts of Interest:** The authors declare no conflicts of interest.



## References

1. Khanam, U.A.; Gao, Z.; Adamko, D.; Kusalik, A.; Rennie, D.C.; Goodridge, D.; Chu, L.; Lawson, J.A. A scoping review of asthma and machine learning. *J. Asthma* **2023**, *60*, 213–226. [\[CrossRef\]](#) [\[PubMed\]](#)
2. Tsang, K.C.H.; Pinnock, H.; Wilson, A.M.; SHAH, S.A. Application of Machine Learning Algorithms for Asthma Management with mHealth: A Clinical Review. *J. Asthma Allergy* **2022**, *15*, 855–873. [\[CrossRef\]](#)
3. Anderson, H.R.; Gupta, R.; Kapetanakis, V.; Asher, M.I.; Clayton, T.; Robertson, C.F.; Strachan, D.P. International correlations between indicators of prevalence, hospital admissions and mortality for asthma in children. *Int. J. Epidemiol.* **2008**, *37*, 573–582. [\[CrossRef\]](#)
4. Ebmeier, S.; Thayabaran, D.; Braithwaite, I.; Bénamara, C.; Weatherall, M.; Beasley, R. Trends in international asthma mortality: Analysis of data from the WHO Mortality Database from 46 countries (1993–2012). *Lancet* **2017**, *390*, 935–945. [\[CrossRef\]](#) [\[PubMed\]](#)
5. Razavi-Termeh, S.V.; Sadeghi-Niaraki, A.; Choi, S.M. Asthma-prone areas modeling using a machine learning model. *Sci. Rep.* **2021**, *11*, 1912. [\[CrossRef\]](#) [\[PubMed\]](#)
6. Hwang, H.; Jang, J.H.; Lee, E.; Park, H.S.; Lee, J.Y. Prediction of the number of asthma patients using environmental factors based on deep learning algorithms. *Respir. Res.* **2023**, *24*, 302. [\[CrossRef\]](#)
7. Chen, Y.; Kong, D.; Fu, J.; Zhang, Y.; Zhao, Y.; Liu, Y.; Chang, Z.; Liu, Y.; Liu, X.; Xu, K.; et al. Associations between ambient temperature and adult asthma hospitalizations in Beijing, China: A time-stratified case-crossover study. *Respir. Res.* **2022**, *23*, 1–12. [\[CrossRef\]](#) [\[PubMed\]](#)
8. Zhou, X.; Sampath, V.; Nadeau, K.C. Effect of air pollution on asthma. *Ann. Allergy Asthma Immunol.* **2024**, *132*, 426–432. [\[CrossRef\]](#)
9. D’Amato, G.; Holgate, S.T.; Pawankar, R.; Ledford, D.K.; Cecchi, L.; Al-Ahmad, M.; Al-Enezi, F.; Al-Muhsen, S.; Ansotegui, I.; Baena-Cagnani, C.E.; et al. Meteorological conditions, climate change, new emerging factors, and asthma and related allergic disorders. A statement of the World Allergy Organization. *World Allergy Organ. J.* **2015**, *8*, 25. [\[CrossRef\]](#) [\[PubMed\]](#)
10. D’Amato, G.; Liccardi, G.; D’Amato, M.; Holgate, S. Environmental risk factors and allergic bronchial asthma. *Clin. Exp. Allergy* **2005**, *35*, 1113–1124. [\[CrossRef\]](#)
11. Menezes, A.M.B.; Wehrmeister, F.C.; Horta, B.; Szwarcwald, C.L.; Vieira, M.L.; Malta, D.C. Prevalence of asthma medical diagnosis among Brazilian adults: National Health Survey, 2013. *Rev. Bras. Epidemiol.* **2015**, *18*, 204–213. [\[CrossRef\]](#) [\[PubMed\]](#)
12. Cardoso, T.A.; Roncada, C.; Silva, E.R.; Pinto, L.A.; Jones, M.H.; Stein, R.T.; Pitrez, P.M. Impacto da asma no Brasil: Análise longitudinal de dados extraídos de um banco de dados governamental brasileiro. *J. Bras. Pneumol.* **2017**, *43*, 163–168. [\[CrossRef\]](#)
13. Pitrez, P.M.; Giavina-Bianchi, P.; Rizzo, J.Â.; Souza-Machado, A.; Garcia, G.F.; Pizzichini, M.M.M. An expert review on breaking barriers in severe asthma in Brazil: Time to act. *Chronic Respir. Dis.* **2021**, *18*, 14799731211028259. [\[CrossRef\]](#)
14. Amorim, J.R.G.; Oliveira, A.M.; Neves, D.; Oliveira, G.P. Associação entre variáveis ambientais e doenças respiratórias (asma e bronquite) em crianças na cidade Macapá-AP no período de 2008 a 2012. *Planeta Amaz. Rev. Int. Direito Ambient. Políticas Públicas* **2013**, *5*, 141–153.
15. Andrade, D.O.; Botelho, C.; Junior, J.L.R.S.; Faria, S.S.; Rabahi, M.F. Sazonalidade Climática E Hospitalizações Em Crianças Menores De Cinco Anos Com Doença Respiratória, Goiânia/GO. *Rev. Bras. Geogr. Médica Saúde* **2015**, *11*, 99–105. [\[CrossRef\]](#)
16. Savian, M.C.B.; Jacobi, L.F.; Zanini, R.R. A relação entre o número de internações por doenças respiratórias e variáveis climáticas em Santa Maria—RS. *Ciência Nat.* **2020**, *42*, e53. [\[CrossRef\]](#)
17. Xavier, J.M.V.; Silva, F.D.S.; Olinda, R.A.; Querino, L.A.L.; Araujo, P.S.B.; Lima, L.F.C.; Sousa, R.S.; Rosado, B.N.C.L. Climate seasonality and lower respiratory tract diseases: A predictive model for pediatric hospitalizations. *Rev. Bras. Enferm.* **2022**, *75*, e20210680. [\[CrossRef\]](#)
18. Marques, C.P.C.; Bloise, R.F.; Lopes, L.B.M.; Godói, L.F.; Souza, P.R.P.; Rosa, I.M.S.; Costa, S.S.; Barros, M.C.; Souza, A.C.L.; Carvalho, B.M.M. Epidemiologia da Asma no Brasil, no período de 2016 a 2020. *Res. Soc. Dev.* **2022**, *11*, e5211828825. [\[CrossRef\]](#)
19. Santo, A.H. Mortalidade relacionada à asma, Brasil, 2000: Um estudo usando causas múltiplas de morte. *Cad. Saúde Pública* **2006**, *22*, 41–52. [\[CrossRef\]](#)
20. Ribeiro-Silva, R.C.; Barreto, M.L.; Ramos, D.; Cruz, A.A.; Oliveira-Campos, M.O.; Malta, D.C. Asthma trend in adolescence in Brazil: Results of the National Adolescent Schoolbased Health Survey (PeNSE 2012–2015). *Rev. Bras. Epidemiol.* **2018**, *21*, e180017. [\[CrossRef\]](#) [\[PubMed\]](#)
21. Jordan, M.I.; Mitchell, T.M. Machine learning: Trends, perspectives, and prospects. *Science* **2015**, *349*, 253–255. [\[CrossRef\]](#)
22. IBGE—Instituto Brasileiro de Geografia e Estatística. *Censo Brasileiro de 2012*; IBGE: Rio de Janeiro, Brazil, 2022.
23. DATASUS, Ministério da Saúde. Sistema de Informação Sobre Internação, 1979–1997. Available online: <http://www.datasus.gov.br/> (accessed on 1 September 2024).
24. Inness, A.; Blechschmidt, A.M.; Bouarar, I.; Chabrilat, S.; Crepulja, M.; Engelen, R.J.; Eskes, H.; Flemming, J.; Gaudel, A.; Hendrick, F.; et al. Data assimilation of satellite-retrieved ozone, carbon monoxide and nitrogen dioxide with ECMWF’s Composition-IFS. *Atmos. Chem. Phys.* **2015**, *15*, 5275–5303. [\[CrossRef\]](#)

25. Flemming, J.; Benedetti, A.; Inness, A.; Engelen, R.J.; Jones, L.; Huijnen, V.; Remy, S.; Parrington, M.; Suttie, M.; Bozzo, A.; et al. The CAMS interim Reanalysis of Carbon Monoxide, Ozone and Aerosol for 2003–2015. *Atmos. Chem. Phys.* **2017**, *17*, 1945–1983. [CrossRef]
26. Inness, A.; Ades, M.; Agustí-Panareda, A.; Barré, J.; Benedictow, A.; Blechschmidt, A.-M.; Dominguez, J.J.; Engelen, R.; Eskes, H.; Flemming, J.; et al. The CAMS reanalysis of atmospheric composition. *Atmos. Chem. Phys.* **2019**, *19*, 3515–3556. [CrossRef]
27. Agustí-Panareda, A.; Barré, J.; Massart, S.; Inness, A.; Absen, I.; Ades, M.; Baier, B.C.; Balsamo, G.; Borsdorff, T.; Bousserez, N.; et al. Technical note: The CAMS greenhouse gas reanalysis from 2003 to 2020. *Atmos. Chem. Phys.* **2023**, *23*, 3829–3859. [CrossRef]
28. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
29. Ho, T.K. Random decision forest. In Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, QC, Canada, 14–16 August 1995; pp. 278–282.
30. Kansara, D.; Singh, R.; Sanghvi, D.; Kanani, P. Improving Accuracy of Real Estate Valuation Using Stacked Regression. *Int. J. Eng. Dev. Res.* **2018**, *6*, 571–577.
31. Chahboun, S.; Maaroufi, M. Principal Component Analysis and Machine Learning Approaches for Photovoltaic Power Prediction: A Comparative Study. *Appl. Sci.* **2021**, *11*, 7943. [CrossRef]
32. Pal, M. Random forest classifier for remote sensing classification. *Int. J. Remote Sens.* **2005**, *26*, 217–222. [CrossRef]
33. Farhangi, F.; Sadeghi-Niaraki, A.; Nahvi, A.; Razavi-Termeh, S.V. Spatial modelling of accidents risk caused by driver drowsiness with data mining algorithms. *Geocarto Int.* **2022**, *37*, 2698–2716. [CrossRef]
34. Chen, T.; Guestrin, C. XGBoost: A scalable tree boosting system. In Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794.
35. Aaron, S.D.; Boulet, L.P.; Reddel, H.K.; Gershon, A.S. Underdiagnosis and Overdiagnosis of Asthma. *Am. J. Respir. Crit. Care Med.* **2018**, *198*, 1012–1020. [CrossRef] [PubMed]
36. Tomita, K.; Yamasaki, A.; Katou, R.; Ikeuchi, T.; Touge, H.; Sano, H.; Tohda, Y. Construction of a Diagnostic Algorithm for Diagnosis of Adult Asthma Using Machine Learning with Random Forest and XGBoost. *Diagnostics* **2023**, *13*, 3069. [CrossRef]
37. Maulud, D.; Abdulazeez, A.M. A Review on Linear Regression Comprehensive in Machine Learning. *J. Appl. Sci. Technol. Trends* **2020**, *1*, 140–147. [CrossRef]
38. Cortes, C.; Vapnik, V. Support-vector networks. *Mach. Learn.* **1995**, *20*, 273–297. [CrossRef]
39. Burges, C.J.C. A Tutorial on Support Vector Machines for Pattern Recognition. *Data Min. Knowl. Discov.* **1998**, *2*, 121–167. [CrossRef]
40. Clarke, S.M.; Griebisch, J.H.; Simpson, T.W. Analysis of Support Vector Regression for Approximation of Complex Engineering Analyses. *J. Mech. Des.* **2005**, *127*, 1077–1087. [CrossRef]
41. Noble, W.S. What is a support vector machine? *Nat. Biotechnol.* **2006**, *24*, 1565–1567. [CrossRef]
42. Alpaydin, E. Voting over Multiple Condensed Nearest Neighbors. *Artif. Intell. Rev.* **1997**, *11*, 115–132. [CrossRef]
43. Shouman, M.; Turner, T.; Stocker, R. Applying k-Nearest Neighbour in Diagnosing Heart Disease Patients. *Int. J. Inf. Educ. Technol.* **2012**, *2*, 220–223. [CrossRef]
44. Botchkarev, A. Evaluating Performance of Regression Machine Learning Models Using Multiple Error Metrics in Azure Machine Learning Studio. Available online: [https://www.academia.edu/36631457/Evaluating\\_performance\\_of\\_regression\\_machine\\_learning\\_models\\_using\\_multiple\\_error\\_metrics\\_in\\_Azure\\_Machine\\_Learning\\_Studio](https://www.academia.edu/36631457/Evaluating_performance_of_regression_machine_learning_models_using_multiple_error_metrics_in_Azure_Machine_Learning_Studio), (accessed on 10 October 2024).
45. Botchkarev, A. A New Typology Design of Performance Metrics to Measure Errors in Machine Learning Regression Algorithms. *Interdiscip. J. Inf. Knowl. Manag.* **2019**, *14*, 45–76. [CrossRef]
46. Li, X.; Zhang, X. A comparative study of statistical and machine learning models on carbon dioxide emissions prediction of China. *Environ. Sci. Pollut. Res.* **2023**, *30*, 117485–117502. [CrossRef] [PubMed]
47. Alizamir, M.; Kim, S.; Kisi, O.; Zounemat-Kermani, M. A comparative study of several machine learning based non-linear regression methods in estimating solar radiation: Case studies of the USA and Turkey regions. *Energy* **2020**, *197*, 117239. [CrossRef]
48. Doreswamy, K.S.H.; Yogesh, K.M.; Gad, I. Forecasting Air Pollution Particulate Matter (PM<sub>2.5</sub>) Using Machine Learning Regression Models. *Procedia Comput. Sci.* **2020**, *171*, 2057–2066. [CrossRef]
49. Chicco, D.; Warrens, M.J.; Jurman, G. The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. *PeerJ Comput. Sci.* **2021**, *7*, e623. [CrossRef]
50. Box, G.E.P.; Pierce, D.A. Distribution of Residual Autocorrelations in Autoregressive-Integrated Moving Average Time Series Models. *J. Am. Stat. Assoc.* **1970**, *65*, 1509–1526. [CrossRef]
51. Brockwell, P.J.; Davis, R.A. *Time Series: Theory and Methods*; Springer: New York, NY, USA, 1991.
52. Hassani, H.; Royer-Carenzi, M.; Mashhad, L.M.; Yarmohammadi, M.; Yeganegi, M.R. Exploring the Depths of the Autocorrelation Function: Its Departure from Normality. *Information* **2024**, *15*, 449. [CrossRef]
53. Achakulwisut, P.; Brauer, M.; Hystad, P.; Anenberg, S.C. Global, national, and urban burdens of paediatric asthma incidence attributable to ambient NO<sub>2</sub> pollution: Estimates from global datasets. *Lancet Planet. Health* **2019**, *3*, e166–e178. [CrossRef]

54. Ding, L.; Zhu, D.; Peng, D.; Zhao, Y. Air pollution and asthma attacks in children: A case e crossover analysis in the city of Chongqing, China. *Environ. Pollut.* **2017**, *220*, 348–353. [[CrossRef](#)]
55. Razavi-Termeh, S.V.; Sadeghi-Niaraki, A.; Choi, S.M. Effects of air pollution in Spatio-temporal modeling of asthma-prone areas using a machine learning model. *Environ. Res.* **2021**, *200*, 111344. [[CrossRef](#)]
56. World Health Organization (WHO). *Policies, Regulations & Legislation Promoting Healthy Housing: A Review*; WHO: Geneva, Switzerland, 2021.
57. Chen, B.-Y.; Chen, C.-H.; Chuang, Y.-C.; Wu, Y.-H.; Pan, S.-C.; Guo, Y.L. Changes in the relationship between childhood asthma and ambient air pollution in Taiwan: Results from a nationwide survey repeated 5 years apart. *Pediatr. Allergy Immunol.* **2019**, *30*, 188–194. [[CrossRef](#)] [[PubMed](#)]
58. Deng, Q.; Deng, L.; Lu, C.; Li, Y.; Norbäck, D. Parental stress and air pollution increase childhood asthma in China. *Environ. Res.* **2018**, *165*, 23–31. [[CrossRef](#)]
59. Hwang, B.-F. Traffic related air pollution as a determinant of asthma among Taiwanese school children. *Thorax* **2005**, *60*, 467–473. [[CrossRef](#)]
60. Laurent, O.; Pedrono, G.; Segala, C.; Filleul, L.; Havard, S.; Deguen, S.; Schillinger, C.; Rivière, E.; Bard, D. Air pollution, asthma attacks, and socioeconomic deprivation: A small-area case-crossover study. *Am. J. Epidemiol.* **2008**, *168*, 58–65. [[CrossRef](#)] [[PubMed](#)]
61. Li, S.; Batterman, S.; Wasilevich, E.; Wahl, R.; Wirth, J.; Su, F.-C.; Mukherjee, B. Association of daily asthma emergency department visits and hospital admissions with ambient air pollutants among the pediatric Medicaid population in Detroit: Time-series and time-stratified case-crossover analyses with threshold effects. *Environ. Res.* **2011**, *111*, 1137–1147. [[CrossRef](#)]
62. Lovinsky-Desir, S.; Acosta, L.M.; Rundle, A.G.; Miller, R.L.; Goldstein, I.F.; Jacobson, J.S.; Chillrud, S.N.; Perzanowski, M.S. Air pollution, urgent asthma medical visits and the modifying effect of neighborhood asthma prevalence. *Pediatr. Res.* **2019**, *85*, 36–42. [[CrossRef](#)] [[PubMed](#)]
63. Samoli, E.; Nastos, P.T.; Paliatsos, A.G.; Katsouyanni, K.; Priftis, K.N. Acute effects of air pollution on pediatric asthma exacerbation: Evidence of association and effect modification. *Environ. Res.* **2011**, *111*, 418–424. [[CrossRef](#)]
64. Santus, P.; Russo, A.; Madonini, E.; Allegra, L.; Blasi, F.; Centanni, S.; Miadonna, A.; Schiraldi, G.; Amaducci, S. How air pollution influences clinical management of respiratory diseases. A case-crossover study in Milan. *Respir. Res.* **2012**, *13*, 95. [[CrossRef](#)]
65. Son, J.-Y.; Lee, J.-T.; Park, Y.H.; Bell, M.L. Short-Term Effects of Air Pollution on Hospital Admissions in Korea. *Epidemiology* **2013**, *24*, 545–554. [[CrossRef](#)]

**Disclaimer/Publisher’s Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.