earth

MDPI

Article

Optimizing Air Pollution Forecasting Across Temporal Scales: A Case Study in Salamanca, Mexico

Francisco-Javier Moreno-Vazquez, Felipe Trujillo-Romero * and Amanda Enriqueta Violante Gavira

División de Ingenierías Campus Irapuato—Salamanca, Universidad de Guanajuato, Guanajuato 36787, Mexico; fj.morenovazquez@ugto.mx (F.-J.M.-V.); amanda@ugto.mx (A.E.V.G.)
* Correspondence: fdj.trujillo@ugto.mx

Abstract: Air pollution forecasting is essential for understanding environmental patterns and mitigating health risks, especially in urban areas. This study investigates the forecasting of criterion pollutants—CO, O_3, SO_2, NO_2, $PM_{2.5}$, and PM_{10}—across multiple temporal frames (hourly, daily, weekly, monthly) in Salamanca, Mexico, utilizing temporal, meteorological, and pollutant data from local monitoring stations. The primary objective is to identify robust models capable of short- and mid-term predictions, despite challenges related to data inconsistencies and missing values. Leveraging the low-code PyCaret framework, a benchmark analysis was conducted to identify the best-performing models for each pollutant. Statistical evaluations, including ANOVA and Tukey HSD tests, were employed to compare model performance across different time frames. The results reveal significant variations in prediction accuracy depending on both the pollutant and temporal windows, with stronger predictive performance observed in the weekly and monthly frames. The research indicates that the incorporation of temporal and environmental variables enhances forecast accuracy and highlights the value of low-code AutoML tools, such as PyCaret, in streamlining model selection and improving overall forecasting efficiency.

Keywords: air pollution; forecast; machine learning; air quality; PyCaret

check for updates

Academic Editor: Carmine Serio

Received: 9 December 2024
Revised: 5 February 2025
Accepted: 5 February 2025
Published: 9 February 2025

Citation: Moreno-Vazquez, F.-J.; Trujillo-Romero, F.; Violante Gavira, A.E. Optimizing Air Pollution Forecasting Across Temporal Scales: A Case Study in Salamanca, Mexico. *Earth* **2025**, *6*, 9. https://doi.org/10.3390/earth6010009

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/).

1. Introduction

At present, in the field of environmental management and care, one of the most prominent areas of focus is air quality. Emphasis is given to the common issue faced by a significant number of Latin American cities that experience the severe effects of atmospheric pollution, which is considered a public health problem [1]. Atmospheric pollutants are substances capable of negatively altering the natural composition of air in a specific geographic area over a defined period [2]. Therefore, the Environmental Protection Agency (EPA) has recommended that governments monitor and regulate these substances under the classification of "criteria pollutants" and "non-criteria pollutants". The former category represents a greater threat to human health and the environment [3], while non-criteria pollutants pose a lower health risk and thus do not require continuous monitoring.

The consumption of fossil fuels, which gives rise to criteria pollutants, causes a deterioration in the quality of life and health of people [4] also affecting productive activities [5] and the environment [6]. In general, the main health effects of air pollution include allergic diseases [7], diabetes mellitus [8], respiratory and cardiovascular diseases, cancer, and premature deaths [9] (See Table 1).

Among the most aggressive pollutants to health are fine particles with a diameter of less than ten microns (PM_{10}) and 2.5 microns ($PM_{2.5}$), ozone (O_3), nitrogen oxides

(NO_2), carbon monoxide (CO), and sulfur dioxide (SO_2), which are generated in highly industrialized cities [7].

On the other side, machine learning algorithms are being used to predict pollution levels, addressing environmental challenges such as increased pollutant concentrations. These algorithms, including Extra Trees regressor, random forest regressor, light gradient boosting machine, and K-nearest neighbor, were evaluated in [10]. Udin et al. [11] used a K-nearest neighbor (kNN) machine learning model to predict $PM_{2.5}$ concentrations in Gazipur City, Bangladesh. The model, based on a dataset from 2012–2018, was evaluated using error metrics like the $MAE, MSE, RMSE, and\ R^2$. The results showed accurate and reliable predictions, potentially contributing to sustainable air pollution prediction strategies to improve air quality. In [12], the authors present a location-invariant air pollution prediction model using a light GBR, random forest, and transfer learning strategy. Experiments show the model outperforms standard models in forecasting air pollution levels, providing valuable information for environmental planning and potentially improving air quality management and decision making across regions, thereby safeguarding the environment and public health.

Narkede et al. [13] proposed two innovative methods to predict the $PM_{2.5}$ concentration in advance that involve developing new indexing methods, the Hourly Relative Mean Index and the Hourly Weighted Index. The methods were used to forecast $PM_{2.5}$ concentrations for an hourly time frame and 48 h ahead. In [14], Arnaut et al. present a novel approach for imputing missing values in air quality data using the random forest (RF) algorithm. The algorithm imputes missing observations in a bi-directional manner for air quality data from the Republic of Serbia, using $PM_{2.5}$ data. The algorithm was evaluated against simple methods, yielding error rates comparable to those of the median imputation method for all periods. In another study, Zareba et al. [15] explored the impact of energy transformation and environmental policy on air pollution in Krakow, Poland, by means of machine learning and artificial intelligence; the authors analyzed historical data and predicted future events using IoT sensors using various ML approaches, including regression models, deep neural networks, and ensemble learning, to predict air pollution. They found that population growth does not negatively affect PM concentrations, and transitioning from coal-based to sustainable energy sources is the primary factor in improving air quality.

In [16] a stacking model called AirStackNet was developped, which uses first-person photos for predicting air quality by means of a neural network of two layers and four regression models to generate metadata from the first layer; this model improved on the air pollution prediction performance by overcoming the bias–variance tradeoff and outperformed baseline and state-of-the-art models, highlighting the importance of addressing air quality in our daily lives. Also, Rad et al. [17] used a convolutional neural network to forecast air pollution concentrations in Tehran, Iran. The implemented system performed well in predicting NO_2, PM_{10}, and $PM_{2.5}$ concentrations, and they found that the temperature, humidity, $PM_{2.5}$, and PM_{10} are the primary factors affecting pollutant concentration variability.

Thanks to the evolution of machine learning, more and more tools like PyCaret are being developed and used by researchers in different areas. For example, Westergaard et al. [18] evaluated three machine learning tools, AutoGluon, Auto-Sklearn, and PyCaret, using diverse datasets for time series analysis and highlighted the ongoing need for the research and development of these tools for different specialized areas. In the same way, Mazahir et al. [19] explored the use of automated machine learning (AutoML) and real-time weather data to improve rain prediction. This approach improved the rain prediction accuracy, with scores ranging from 80.60 to 86.09% across different algorithms. The authors of [20] improved traditional exposure assessment methods for $PM_{2.5}$ prediction by incorpo-

rating measurements from a low-cost sensor network, regulatory monitors, an automated machine learning modeling framework, and big data on human mobility.

Finally, we mention the work of Quintana et al. [21], which used PyCaret to analyze historical data on monthly electricity consumption for the Cuajone Mining Unit of the Minera Southern Peru Copper Corporation; they evaluated 27 machine learning models, focusing on Exponential Smoothing, AdaBoost with Conditional Deseasonalization and Detrending, and ETS (Error Trend Seasonality). The models performed well, with Exponential Smoothing showing the best performance with a $MASE$ of 0.8359, MAE of 4012.24, and $RMSE$ of 5922.63.

Table 1. Criteria for pollutants, sources, and health effects.

Criteria Pollutants	Source	Health Effects
Particulate matter (PM_{10} and $PM_{2.5}$)	Comes from the suspension or re-suspension of dust, soil, or other materials from roads, agriculture, mining, combustion processes of motor vehicles, energy generation, and industrial processes, as well as wind or volcanic storms.	Ultrafine particles can penetrate the deepest part of the respiratory tract, causing respiratory problems, exacerbating asthma, and damaging lung function. Diesel-derived particles may pose a cancer risk.
Nitrogen dioxide (NO_2)	Results from high-temperature combustion (motor vehicles, industrial boilers, and thermoelectric plants).	
Tropospheric Ozone (O_3)	Forms in the atmosphere through reactions of its precursors (NOx and VOCs) in the presence of sunlight.	Associated with reduced lung function in individuals engaged in light to heavy exercise.
Sulfur Dioxide (SO_2)	Comes from the burning of fuels (coal with high-sulfur content and oil) from electric companies, industrial processes, and metal extraction from minerals.	Emissions from vehicular sources cause changes in lung function in asthmatics and exacerbate respiratory symptoms in sensitive individuals. It causes coughing, mucus secretion, and chronic bronchitis. It increases the propensity of sensitive individuals to contract respiratory infections.
Carbon Monoxide (CO)	Results from the incomplete combustion of natural gas, coal, or wood. One potential source is vehicular traffic.	Inhibits the blood's ability to carry oxygen to organs and tissues. People with chronic heart disease may experience chest pains when levels are high. At very high levels, it impairs vision, manual dexterity, and learning capacity and can lead to death.

Recent events related to environmental and health deterioration have stimulated research in this field, particularly in cities with industrial development, where there is an increase in the production of polluting gases from industry and motor vehicles [22], such as the case of Salamanca, Guanajuato, Mexico. The city of Salamanca belongs to a region known as the "Bajio", which includes the states of Guanajuato, Queretaro, Aguascalientes, and Jalisco. This region is considered a commercial hub between Mexico, the United States, and Central America, which has triggered the spatial growth of the city. The economic dynamics revolve mainly around the industrial sector. Salamanca is one of 46 municipalities in the state of Guanajuato, with the city of the same name as the municipal seat. It has a territorial area of 756.54 km² and is located at 20°34′13″ north latitude and 101°11′50″ west

longitude, at 1721 m above sea level. In this city, the main sources of pollutant emissions are the Antonio M. Amor Refinery and the Salamanca Thermoelectric Plant [23]. To monitor air quality, the municipality has three atmospheric monitoring stations: Cruz Roja, Nativitas, and DIF. These stations are equipped with automatic technology, including particle samplers and gas analyzers. Additionally, they are capable of measuring meteorological data such as temperature, relative humidity, atmospheric pressure, and wind speed and direction. This enables accurate assessments of air quality. This is because environmental authorities must have accurate information that is translated into terms of colors to communicate to the population 24 h a day, 365 days a year.

In addition to direct pollutant measurements, modeling is employed, offering the advantage of reducing equipment operation and maintenance costs. The numerical model is based on the dispersion of air pollutants and simulates dispersion and transport mechanisms using emission source data and information on atmospheric chemical reactions. Advanced statistical models, such as learning methodologies, are also applied [24]. In this context, this study's objective was to compare the different machine learning algorithms provided for an automated tool for predicting concentrations of criteria pollutants at the three monitoring stations in Salamanca.

2. Materials and Methods

This study examined machine learning models for air quality forecasting in Salamanca, Mexico, an urban area with significant industrial activity and consistent air quality monitoring. Data were collected from a monitoring station, including air pollutant concentrations and meteorological variables. The data were preprocessed to address inconsistencies, outliers, and numerical features. Categorical data and temporal features were encoded for machine learning integration. Automated feature selection was performed to identify relevant predictors for air quality forecasting. The experimental setup was designed to systematically compare multiple models under consistent conditions, using Pycaret's low-code interface. Models evaluated included decision trees, random forests, gradient boosting, and ensemble techniques. K-fold cross-validation was applied for robustness. Performance metrics like the $RMSE$, MAE, and R^2 were used to evaluate the models. The results were then thoroughly evaluated to determine the best-performing models, with visualization techniques like error distribution plots and prediction vs. actual scatter plots used to interpret the results. The key stages in the process are described in the next paragraphs.

2.1. Study Area

The study area for this research focused on the city of Salamanca, located in the state of Guanajuato, Mexico. The city is monitored by three key air quality stations, Cruz Roja, Nativitas, and DIF, as shown in Figure 1. These stations provide continuous monitoring of atmospheric pollutants, as shown in Table 2, which details their geographic locations.

Table 2. Statistical descriptions of the temporal variables.

Station Name	Identifier	Latitude	Longitude
Cruz Roja	09	20.5783	−101.1989
Nativitas	10	20.5649	−101.1885
DIF	11	20.5583	−101.2051

Salamanca is a municipality with a growing industrial base that significantly impacts air quality. In recent decades, the city has experienced rapid industrial expansion, driven primarily by the establishment of several large-scale facilities. Salamanca hosts the largest Pemex oil refinery in Mexico, which is also the 11th largest globally in terms of crude oil

processing capacity [25]. Additionally, the CFE Thermoelectric power plant contributes to the industrial landscape and provides substantial local employment. In 2014, Mazda opened its North American manufacturing plant just outside the city, further enhancing the industrial cluster in the region [26].

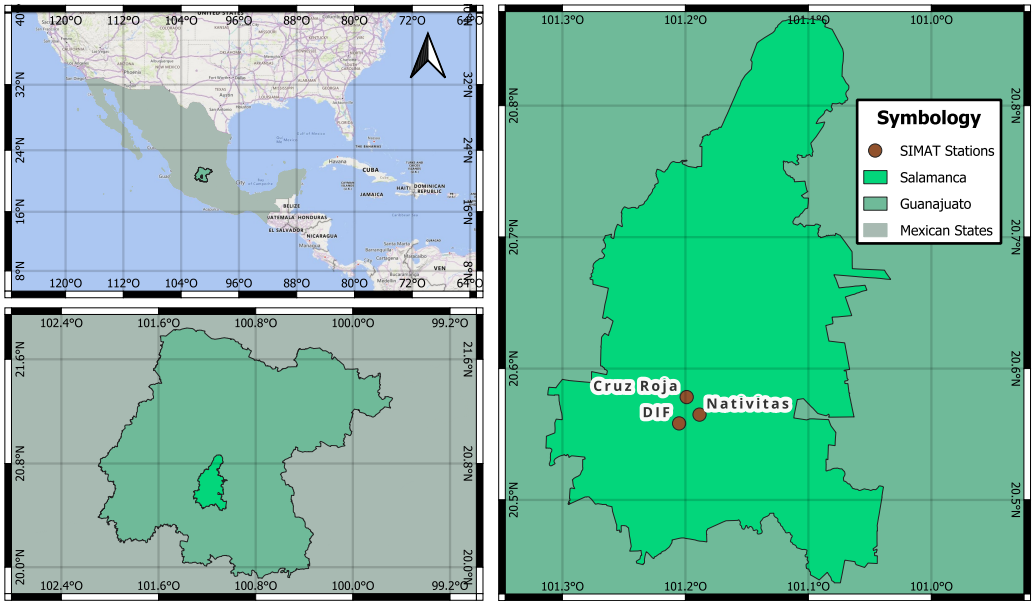

Figure 1. Locations of studied stations. From top to bottom and left to right: Map of Mexico focusing on Guanajuato State, map of Guanajuato State focusing on Salamanca City, map of Salamanca City focusing on SIMAT Stations, plotted by QGIS.

This concentration of industrial activities, combined with increasing urbanization, makes Salamanca a critical area for studying air pollution and its effects. The city's industrial emissions, notably from the oil refinery and thermoelectric plant, have heightened the need for continuous environmental monitoring, making it an ideal location for this research.

2.2. Data Collection

The dataset employed in this study consisted of hourly measurements collected from the SMAOT (Secretaría de Medio Ambiente y Ordenamiento Territorial) monitoring stations [27], spanning from 1 January 2014, to 30 November 2023. After an extensive preprocessing phase—comprising data cleaning, imputation, and threshold-based filtering—the final dataset contained 10,742 daily data points. These data represented air quality and meteorological variables gathered from the three monitoring stations listed in Table 2.

The data collected include both temporal and environmental variables, along with a range of air pollutants, making a comprehensive dataset for studying air quality dynamics in Salamanca. The list of variables used in this study is detailed in Table 3.

Temporal variables are critical for understanding how air quality and environmental conditions evolve over different time scales. By analyzing data on an hourly, daily, weekly, and monthly basis, we gain insights into both short-term fluctuations and long-term trends [28]. The *YEAR* variable provides a general sense of yearly trends, helping to identify long-term shifts in environmental conditions, possibly due to regulatory changes or climate factors. The *MONTH*, *WEEK*, and *DAY* variables allow for a more granular analysis, highlighting seasonal variations and patterns within the year, such as how air quality may change during certain months or days of the week. The *SEASON* variable further emphasizes the role of climate and seasonal factors in shaping air quality, capturing distinct periods (e.g., summer, winter) that are known to influence both meteorological and pollutant levels.

Table 3. List of variables used in this study.

Parameter	Dataset Variable	Data Type
Year	YEAR	Temporal
Month	MONTH	Temporal
Season	WEEK	Temporal
Week	SEASON	Temporal
Month Day	DAY	Temporal
Wind Speed	WS	Environmental
Wind Direction	WD	Environmental
Temperature	TEMP	Environmental
Relative Humidity	HR	Environmental
Barometric Pressure	PBAR	Environmental
Solar Radiation	RADSOL	Environmental
Carbon Monoxide	CO	Pollutant
Ozone	O_3	Pollutant
Sulfur Dioxide	SO_2	Pollutant
Nitrogen Dioxide	NO_2	Pollutant
Particulate Matter 2.5	$PM_{2.5}$	Pollutant
Particulate Matter 10	PM_{10}	Pollutant

Environmental variables such as wind speed (WS), wind direction (WD), temperature (TEMP), relative humidity (HR), barometric pressure (PBAR), and solar radiation (RADSOL) play a pivotal role in determining air quality [29]. Changes in these variables influence the dispersion, concentration, and formation of pollutants in the atmosphere. For example, wind patterns (both speed and direction) dictate the movement of pollutants, while temperature and humidity can affect chemical reactions that either produce or dissipate certain gases, such as ozone. By analyzing these variables at multiple temporal scales, we can observe how environmental conditions shift both within short periods (hourly/daily) and over longer-term trends (weekly/monthly). Table 4 provides a comprehensive statistical description of the environmental variables, enabling a detailed analysis of their behavior and variability over different temporal frames.

Pollutant variables such as particulate matter (PM_{10}, $PM_{2.5}$), sulfur dioxide (SO_2), ozone (O_3), carbon monoxide (CO), and nitrogen dioxide (NO_2) are central to this study as they directly reflect the air quality and its impact on public health [30]. Each pollutant reacts differently to environmental conditions and time frames, with some showing more pronounced daily fluctuations (e.g., CO, which is strongly tied to traffic patterns) and others exhibiting seasonal trends (e.g., O_3, which peaks in warmer months). Analyzing these pollutants across different temporal frames is crucial for identifying trends, hotspots, and outliers that could inform air quality management and public health policies. Table 5 provides a comprehensive statistical description of the pollutan variables, enabling a detailed analysis of their behavior and variability over different temporal frames.

Table 4. Statistical descriptions of the environmental variables.

Time Frame	Parameter	WS	WD	TEMP	HR	PBAR	RADSOL
Hourly	count	254,880.00	254,514.00	254,494.00	222,780.00	177,999.00	79,147.00
	mean	2.69	157.89	20.92	49.10	623.49	233.81
	std	1.52	79.94	5.51	21.49	2.80	343.76
	min	0.01	0.02	0.00	0.50	525.13	0.00
	max	28.87	360.00	44.80	99.96	635.61	1211.04

Table 4. *Cont.*

Time Frame	Parameter	WS	WD	TEMP	HR	PBAR	RADSOL
Daily	count	10,672.00	10,658.00	10,671.00	9350.00	7464.00	3406.00
	mean	2.69	157.95	20.93	49.12	623.49	226.93
	std	1.15	56.07	3.05	14.87	2.54	138.77
	min	0.48	0.69	8.15	1.63	599.09	0.08
	max	21.47	278.96	38.97	98.26	634.03	525.53
Weekly	count	1543.00	1538.00	1542.00	1356.00	1078.00	491.00
	mean	2.69	158.02	20.90	49.19	623.49	226.01
	std	1.00	38.90	2.82	13.21	2.30	136.14
	min	0.73	1.17	12.16	18.63	599.54	0.65
	max	19.78	245.96	31.72	82.12	632.47	515.29
Monthly	count	352.00	352.00	352.00	308.00	246.00	112.00
	mean	2.69	158.22	20.92	49.05	623.48	227.17
	std	0.90	28.59	2.59	11.91	1.96	135.00
	min	1.21	1.50	12.93	27.05	605.85	0.73
	max	18.03	253.33	27.70	72.09	631.49	503.75

2.3. Data Preprocessing

Data preprocessing is a critical step in ensuring the reliability and accuracy of the analysis. Given the raw format and inconsistencies within the historical environmental dataset, several automated steps are required to clean and standardize the data for subsequent modeling.

The data originate from a historical repository in CSV format, including pollutant concentrations and meteorological variables from several monitoring stations. Mobile stations were excluded to maintain consistency and comparability, whereas stationary stations were retained for analysis. The CSV files were renamed and standardized to improve traceability and reduce processing errors. The inconsistency in the initial row was resolved with an automatic header detection mechanism. The dataset's column names were made consistent across all files to maintain homogeneity, irrespective of the collection year. This procedure enabled further analysis and maintained the data structure's uniformity throughout the dataset. The dataset included year and month metadata, providing temporal context for time series research. Missing values were replaced with NaN values, and the method selected to address them involved the elimination of rows with missing data, as the major objective of this research was to analyze the data in their purest form to notice the interactions between them. Additional temporal features like seasons and week numbers were generated, providing temporal granularity for capturing seasonal and weekly pollutant concentration patterns.

The data were consolidated into a unified dataset, including hourly pollutant concentrations and meteorological data from various monitoring stations. These data were then aggregated into a daily dataset for each station, capturing broader air quality trends. Weekly averages were computed from the daily dataset, while monthly averages were generated for seasonal patterns and the evolution of air quality over time. Seasonal datasets were derived from monthly data, grouping data into distinct meteorological seasons for each station. Yearly datasets were created for long-term trend analysis and to obtain insight into the pollution control measures' effectiveness.

Table 5. Statistical description of the pollutant variables.

Time Frame	Parameter	PM_{10}	SO_2	O_3	CO	NO_2	$PM_{2.5}$
Hourly	count	253,054.00	251,434.00	241,101.00	214,910.00	205,665.00	106,321.00
	mean	47.72	9.33	15.38	0.80	7.62	21.50
	std	36.92	16.68	17.98	0.78	8.81	14.66
	min	0.00	0.00	−0.00	−0.01	0.00	0.00
	max	981.61	362.15	137.50	10.40	119.09	416.00
Daily	count	10,700.00	10,651.00	10,275.00	9179.00	8825.00	4653.00
	mean	47.63	9.31	15.35	0.79	7.53	21.37
	std	22.88	11.62	13.60	0.70	7.04	9.68
	min	1.00	0.00	0.00	0.00	0.00	3.10
	max	243.12	102.92	66.79	4.57	51.89	121.38
Weekly	count	1546.00	1543.00	1498.00	1349.00	1300.00	682.00
	mean	47.84	9.27	15.20	0.79	7.48	21.52
	std	19.23	10.09	13.14	0.69	6.64	7.78
	min	10.72	0.00	0.01	0.03	0.00	7.53
	max	133.12	53.47	56.00	3.99	35.45	61.94
Monthly	count	353.00	353.00	347.00	318.00	308.00	159.00
	mean	47.71	9.29	15.25	0.78	7.46	21.39
	std	17.07	9.33	12.88	0.66	6.45	6.14
	min	12.92	0.00	0.02	0.07	0.00	8.63
	max	100.26	40.80	46.33	3.96	27.38	38.77

2.4. Feature Selection

Feature selection is a critical yet challenging step in machine learning, especially for the predictive modeling of complex systems like air quality forecasting [31]. The precision and applicability of machine learning models significantly rely on the identification and selection of the most pertinent features. This paper proposes an automated feature selection technique aimed at systematically eliminating less significant features while preserving those with the highest predictive capability.

Let x_i denote the ith feature in the dataset, and T represent the selection threshold. We define an indicator function $\mathcal{I}(x_i, T)$ that selects a feature if its correlation value with the target variable exceeds or equals the threshold T. The indicator function is formalized in Equation (1).

$$\mathcal{I}(x_i, T) = \begin{cases} 1 & \text{if } correlation(x_i) \geq T, \\ 0 & \text{if } correlation(x_i) < T, \end{cases} \tag{1}$$

Based on this indicator function, the set of most important features, $\mathcal{F}_{\text{important}}$, can be defined as follows:

$$\mathcal{F}_{\text{important}} = \{x_i \mid I(x_i) = 1\} \tag{2}$$

Although the feature selection process is automated, the choice of the threshold T is determined through a comprehensive correlation analysis. For the hourly and daily datasets, correlation heatmaps are presented in Figures 2 and 3, respectively. The threshold used for these datasets was set at $T = 0.1$, meaning only features with a correlation greater than or equal to 0.1 were retained.

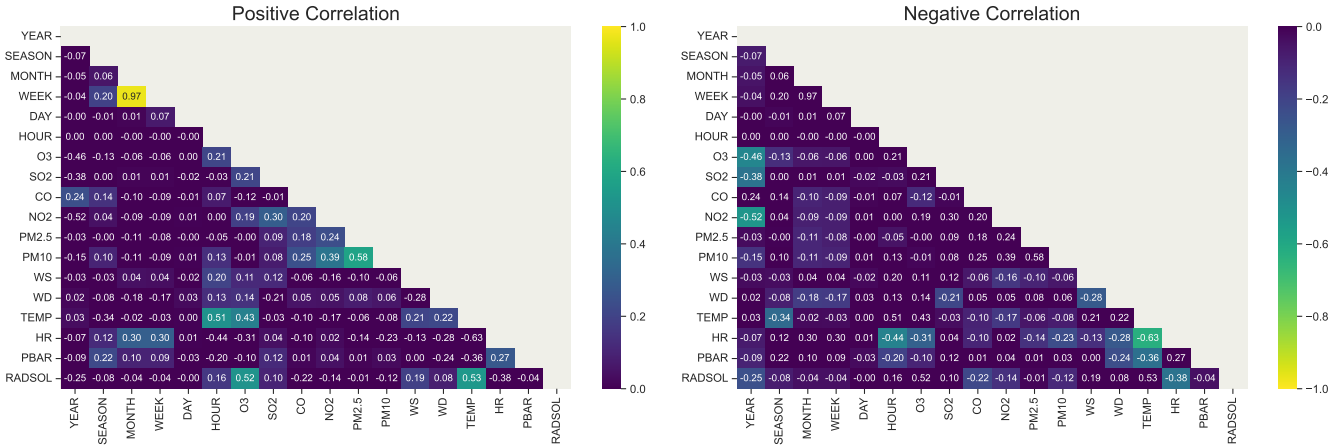

Figure 2. Correlation heatmap for the hourly dataset.

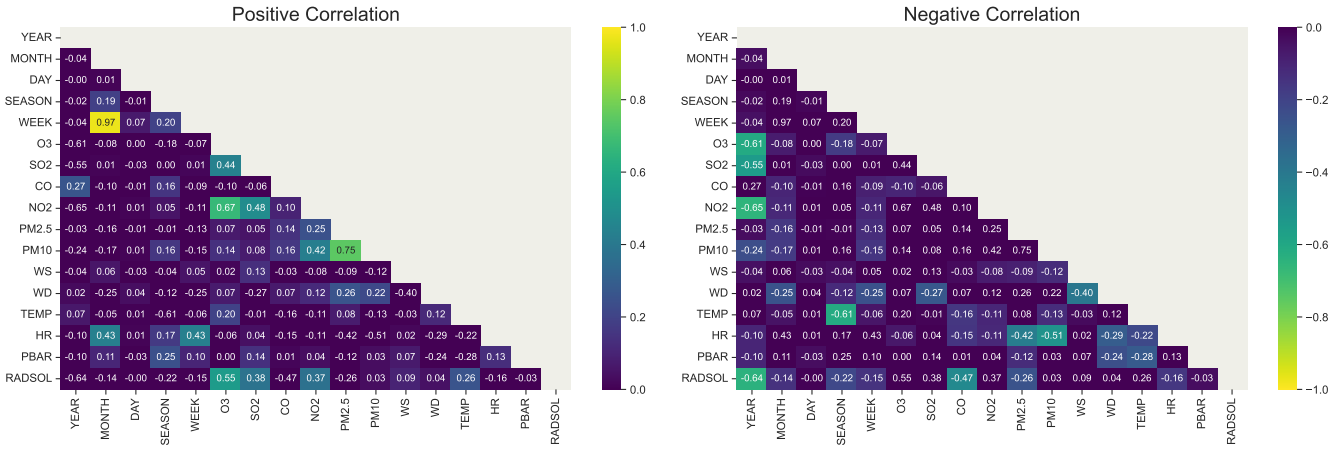

Figure 3. Correlation heatmap for the daily dataset.

For the weekly and monthly datasets, the correlation matrices are shown in Figures 4 and 5, respectively. A slightly higher threshold of $T = 0.14$ was applied for these datasets due to their larger time aggregations, ensuring that only the most relevant features were selected for model training.

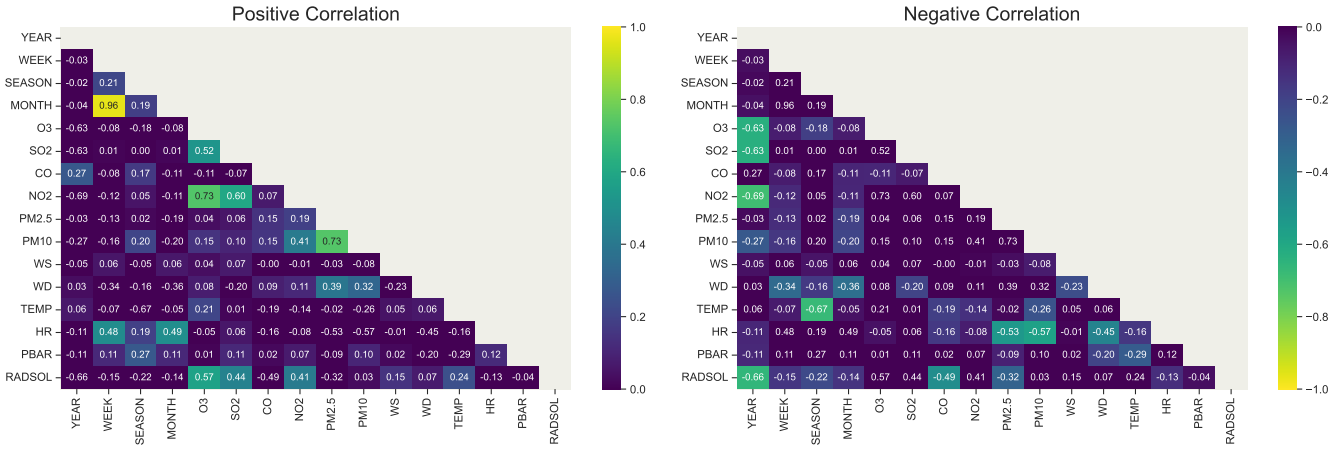

Figure 4. Correlation heatmap for the weekly dataset.

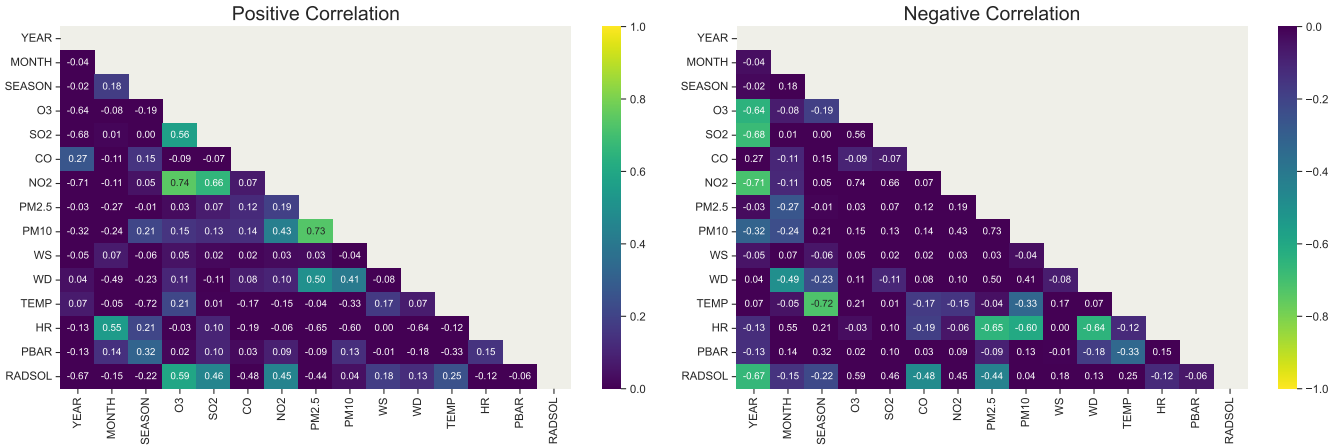

Figure 5. Correlation heatmap for the monthly dataset.

This automatic feature selection method guarantees the retention of the most informative features, enhancing the model's predictive performance while diminishing the dataset's complexity. The selection thresholds were updated dynamically according to the dataset's granularity, enabling effective model training across various temporal resolutions.

2.5. Experimental Setup

After detailing the data characteristics and feature selection process, the next section provides a comprehensive explanation of the experimental procedure followed in this study. It outlines the steps and configurations used to conduct the analysis, including data preparation, model setup, and evaluation methods employed to ensure robust and reproducible results.

3. AutoML Implementation

PyCaret is an open source, low-code machine learning library designed to automate end-to-end machine learning workflows, significantly accelerating the experimental process. By providing an intuitive interface, PyCaret allows users to replace extensive coding efforts with just a few lines, making the experimentation process both faster and more efficient [32].

Unlike many other open source libraries, PyCaret serves as a high-level wrapper around popular machine learning frameworks such as scikit-learn, XGBoost, LightGBM, and CatBoost, among others. It simplifies model selection, training, and optimization through its seamless integration with tools like Optuna [33] and Hyperopt [34], offering an efficient solution for both novices and experienced users [21].

3.1. Hyperparameters

Hyperparameters play a pivotal role in the optimization process of machine learning algorithms, as they directly influence model performance and computational efficiency. Unlike model parameters, which are learned during training, hyperparameters are predefined and control the learning process itself. The appropriate selection and tuning of hyperparameters can drastically improve the model accuracy, convergence speed, and generalization ability, making it a crucial step in any machine learning pipeline [35].

There are several approaches to hyperparameter tuning, ranging from manual search methods, such as trial-and-error, to automated techniques, such as grid search, random search, and more advanced methods like Bayesian optimization and genetic algorithms. PyCaret simplifies this process by employing a randomized grid search for hyperparameter tuning. Randomized grid search offers a more efficient alternative to traditional grid search by sampling a subset of the possible hyperparameter combinations, thereby reduc-

ing the computational cost while still ensuring a high likelihood of finding an optimal configuration [36].

In PyCaret, the hyperparameter search space is defined automatically based on the chosen model, and the random search mechanism explores this space to identify the combination that yields the best performance according to the specified evaluation metric. This automated approach not only streamlines the tuning process but also allows for easy experimentation with various models, enabling users to quickly iterate and refine their machine learning pipelines.

3.2. PyCaret Experiment

In Listing 1, a code snippet from the official PyCaret documentation illustrates the simplicity of the setup for a regression task.

Listing 1. Python code example provided by PyCaret documentation [37].

```
# load sample dataset
from pycaret.datasets import get_data
data = get_data('insurance')
# setup experiment
from pycaret.regression import *
s = setup(data, target = 'charges', session_id = 123)
# get best model best = s.compare_models()
```

In our specific case, the experiment followed a similar setup, with the key difference being the use of our custom datasets derived from the subsystems generated during the automated feature selection process. Each experiment targeted a specific pollutant (*CO*, O_3, SO_2, NO_2, $PM_{2.5}$, PM_{10}) for prediction, as detailed in Table 6. The *setup* function in PyCaret was tailored for each dataset, and we systematically trained, tested, and compared models to identify the best-performing algorithms.

Table 6. Experiment setup for each pollutant in the time frame datasets.

Time Frame	Pollutant	Original Data Shape	Train Set Shape	Test Set Shape	Features
Hourly	CO	5691	4553	1138	8
	O_3	39,029	31,223	7806	11
	SO_2	140,786	11,269	28,157	6
	NO_2	5411	4324	1082	9
	$PM_{2.5}$	80,665	64,532	16,133	6
	PM_{10}	5862	4690	1172	8
Daily	CO	408	326	82	9
	O_3	1813	1450	363	8
	SO_2	1955	1564	391	7
	NO_2	353	282	71	11
	$PM_{2.5}$	298	238	60	9
	PM_{10}	3595	287,612	719	12
Weekly	CO	67	54	13	7
	O_3	369	297	74	7
	SO_2	370	296	74	5
	NO_2	59	47	12	6
	$PM_{2.5}$	50	40	10	7
	PM_{10}	570	456	114	11

Table 6. *Cont.*

Time Frame	Pollutant	Original Data Shape	Train Set Shape	Test Set Shape	Features
Monthly	CO	91	73	18	5
	O_3	90	72	18	7
	SO_2	90	72	18	4
	NO_2	16	13	3	7
	$PM_{2.5}$	16	13	3	6
	PM_{10}	144	115	29	9

3.3. Evaluation Metrics

As part of the evaluation process, the models were tested using the testing dataset to assess their predictive performance. The key metrics used to evaluate the models included the mean absolute error (MAE), mean squared error (MSE), root mean squared error ($RMSE$), coefficient of determination (R^2), root mean squared log error ($RMSLE$), and mean absolute percentage error ($MAPE$). These metrics provide insights into the accuracy, precision, and goodness of fit of the models, enabling a comprehensive assessment of their performance. The equations for calculating these metrics are as follows:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{3}$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{4}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{5}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{6}$$

$$RMSLE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\log(y_i + 1) - log(\hat{y}_i + 1))^2} \tag{7}$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{|y_i - \hat{y}_i|}{y_i} \tag{8}$$

where n is the number of data points, y_i is the actual value of the output variable, $\hat{y}_i$ is the predicted value, and $\bar{y}$ is the mean of the actual values.

4. Results

This section presents the performance results of the selected machine learning models for predicting six key air pollutants: carbon monoxide (CO), ozone (O_3), sulfur dioxide (SO_2), nitrogen dioxide (NO_2), particulate matter smaller than 2.5 microns ($PM_{2.5}$), and particulate matter smaller than 10 microns (PM_{10}). These pollutants were predicted over four temporal resolutions: hourly, daily, weekly, and monthly. The primary aim of this analysis was to evaluate the robustness and generalization capabilities of the models over short- and mid-term prediction windows.

4.1. Overall Results

Table 7 summarizes the best-performing models and their respective error metrics (MAE, MSE, $RMSE$, R^2, $RMSLE$, $MAPE$) for each pollutant and time frame, providing a

comprehensive overview of the experimental results conducted using the PyCaret benchmarking environment.

Table 7. Selected models metrics for each pollutant.

Pollutant	Time Frame	Regressor Model	*MAE*	*MSE*	*RMSE*	R^2	*RMSLE*	*MAPE*
CO	Hourly	Extra Trees	0.204	0.085	0.292	0.805	0.130	0.314
	Daily	Extra Trees	0.116	0.030	0.174	0.894	0.083	0.140
	Weekly	Extra Trees	0.259	0.119	0.345	0.481	0.222	0.972
	Monthly	CatBoost	0.178	0.070	0.265	0.711	0.157	0.545
O_3	Hourly	Extra Trees	2.888	27.899	5.282	0.910	0.293	0.369
	Daily	Extra Trees	2.153	14.098	3.755	0.902	0.187	0.180
	Weekly	Extra Trees	1.810	8.325	2.885	0.932	0.132	0.150
	Monthly	CatBoost	2.598	16.463	4.058	0.907	0.200	2.793
SO_2	Hourly	Extra Trees	6.150	174.908	13.225	0.419	0.710	1.825
	Daily	Extra Trees	2.331	13.936	3.733	0.761	0.351	0.439
	Weekly	Extra Trees	1.533	6.423	2.534	0.868	0.242	0.276
	Monthly	Extreme Gradient Boosting	1.550	5.374	2.318	0.916	0.219	0.322
NO_2	Hourly	CatBoost	0.003	0.000	0.005	0.906	0.005	0.270
	Daily	Extra Trees	0.002	0.000	0.003	0.893	0.003	0.186
	Weekly	AdaBoost	0.004	0.000	0.006	0.039	0.006	0.315
	Monthly	Linear	0.004	0.000	0.005	−2.006	0.005	0.181
$PM_{2.5}$	Hourly	CatBoost	6.278	82.394	9.077	0.595	0.428	0.468
	Daily	Extra Trees	4.484	93.237	9.656	0.579	0.264	0.231
	Weekly	CatBoost	2.246	10.471	3.236	0.744	0.140	0.110
	Monthly	Linear	5.039	38.330	6.191	−0.458	0.237	0.240
PM_{10}	Hourly	Extra Trees	13.016	568.540	23.844	0.648	0.478	0.594
	Daily	CatBoost	5.566	64.317	8.020	0.858	0.223	0.236
	Weekly	CatBoost	4.619	44.241	6.651	0.876	0.138	0.106
	Monthly	CatBoost	7.444	110.103	10.493	0.726	0.216	0.176

It is important to note that the dataset used for training the models contained missing values, which were addressed using imputation techniques. However, the imputation process introduced data inconsistencies, particularly when the input features exhibited strong correlations. These inconsistencies were found to limit data availability, which negatively impacted model performance for certain pollutants and temporal resolutions. Additionally, the variability in model performance between time frames underscores the complexity of the pollutants' behavior across different temporal scales.

The following subsections will detail the performance of the models for each pollutant, focusing on key performance metrics such as the coefficient of determination (R^2), mean absolute error (MAE), and root mean square error ($RMSE$). These metrics provide insights into the models' ability to generalize, account for temporal dynamics, and handle the inherent variability of the pollutants across different time frames.

In Table 7, the Extra Trees regressor showed strong performance in hourly and daily prediction tasks but decreased performance in weekly predictions due to data inconsistencies and reduced feature availability. Meanwhile, CatBoost outperformed Extra Trees for monthly predictions, albeit with a slightly lower R^2 of 0.711, suggesting that it may be better suited for capturing longer-term temporal dynamics of CO. Figure 6 clearly shows the diminishing number of data points for the weekly and monthly predictions, which likely contributed to the observed performance degradation.

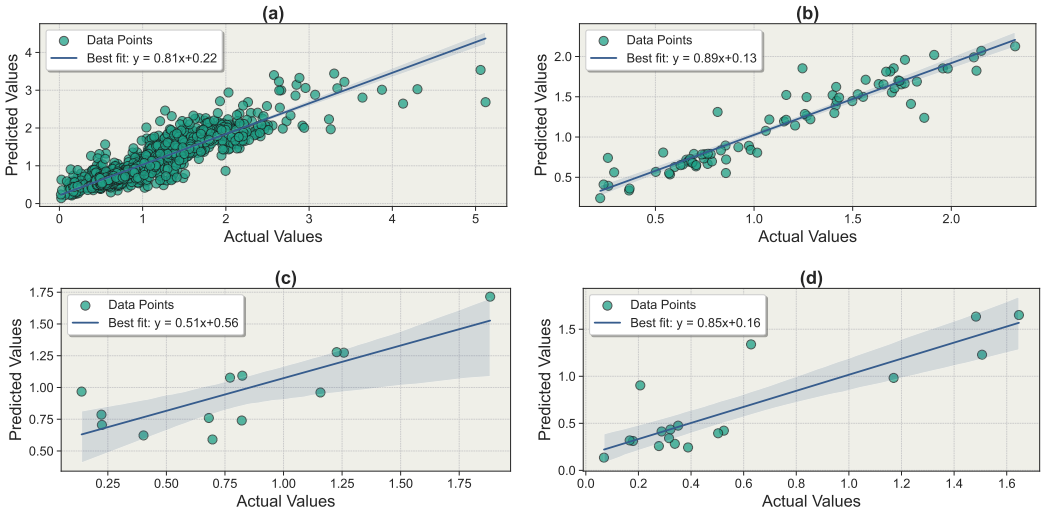

Figure 6. Regression plots of *CO* pollutant for (**a**) hourly, (**b**) daily, (**c**) weekly, and (**d**) monthly time frames.

Moreover, the weekly predictions for O_3 are similar to *CO*, with the Extra Trees regressor showing the best performance across most temporal frames, and CatBoost slightly outperforms it in the monthly frame. The coefficient of determination (R^2) remains close to 0.9 for all predictions, indicating minimal error differences across time frames. The O_3 dataset maintains robust predictive performance until the weekly time frame. Figure 7 presents the regression plots for O_3, highlighting a consistent predictive pattern with slopes exceeding 0.8 across all time frames.

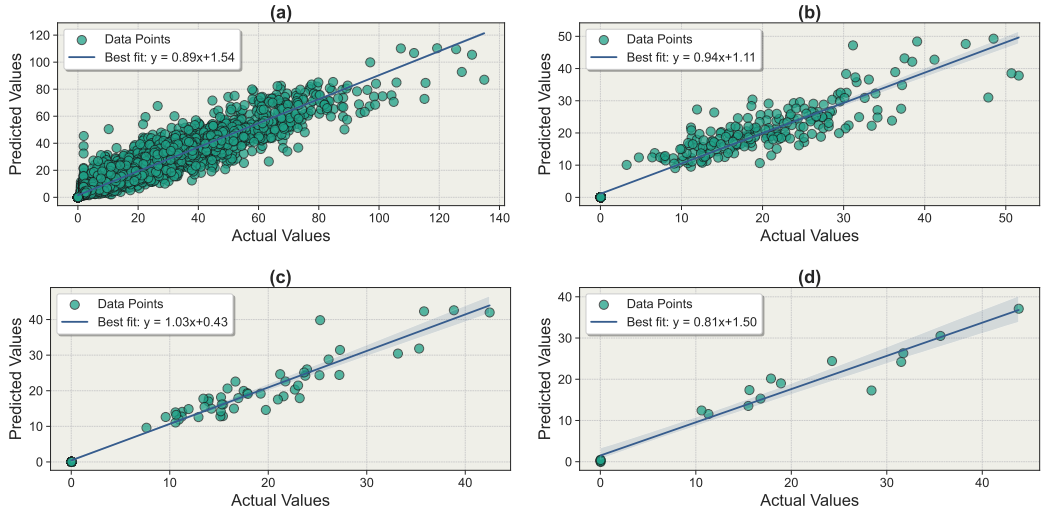

Figure 7. Regression plots of O_3 pollutant for (**a**) hourly, (**b**) daily, (**c**) weekly, and (**d**) monthly time frames.

The Extra Trees regressor's predictions for SO_2 are challenging due to data inconsistencies. Short-term predictions, particularly hourly ones, show poor performance, indicating difficulty in capturing the pollutant's high variability. However, as the prediction window increases, model performance improves. Both Extra Trees and extreme gradient boosting (XGBoost) show strong predictive capability for weekly and monthly forecasts, attributed to data limitations, especially for monthly data. Figure 8 shows the regression plots for SO_2 predictions across different time frames.

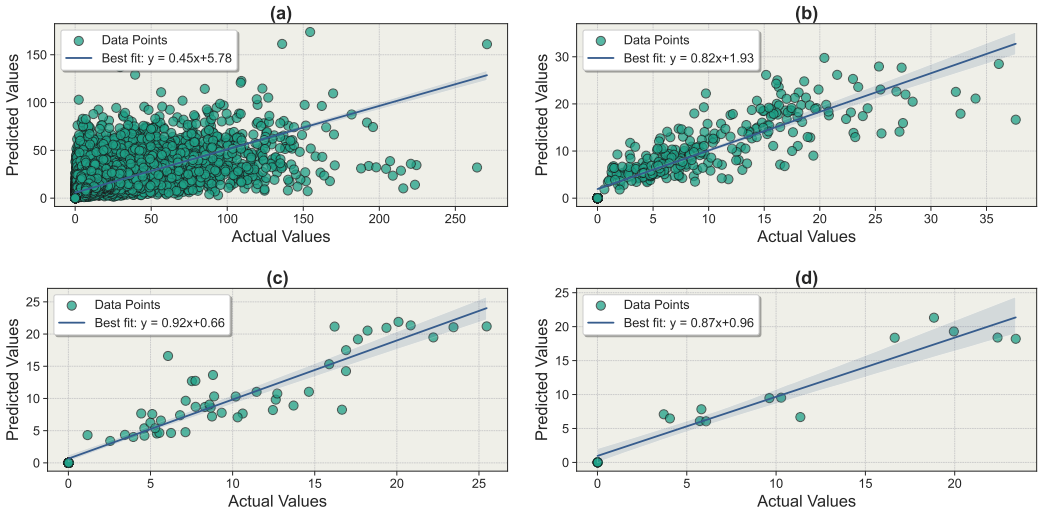

Figure 8. Regression plots for SO_2 predictions across different time frames: (**a**) hourly, (**b**) daily, (**c**) weekly, and (**d**) monthly.

For NO_2 predictions, CatBoost performs best for hourly predictions, with a high R^2 value of 0.906. Extra Trees performs well for daily predictions, with an R^2 of 0.893. However, AdaBoost struggles with weekly and monthly predictions, with a low R^2 of 0.039 and a negative R^2 of -2.006. This suggests that a different modeling approach may be needed for effective monthly NO_2 predictions. Figure 9 illustrates the regression plots for NO_2 predictions across different time frames.

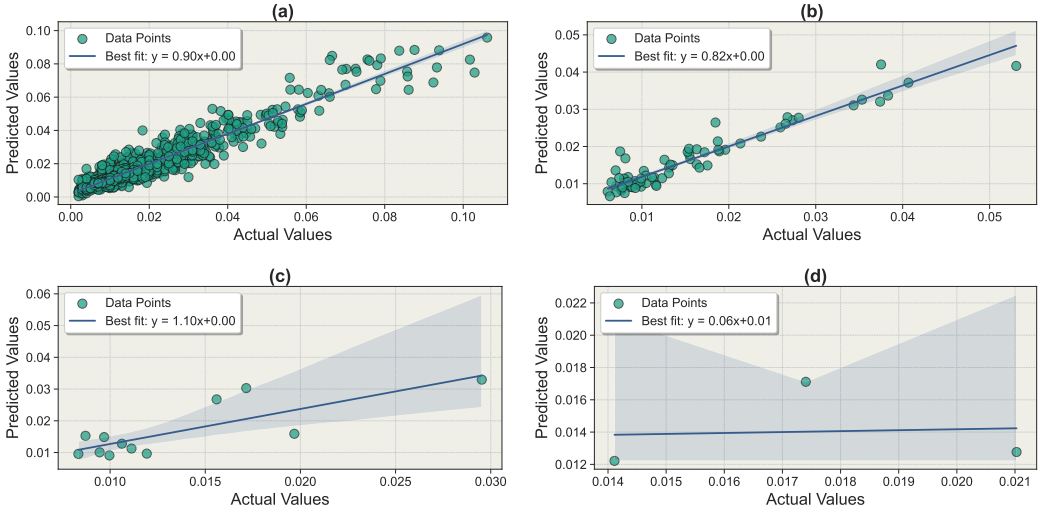

Figure 9. Regression plots of NO_2 pollutant for (**a**) hourly, (**b**) daily, (**c**) weekly, and (**d**) monthly time frames.

In the case of predicting $PM_{2.5}$ levels, CatBoost and Extra Trees are effective models. They perform well for hourly and daily predictions, capturing short-term variations. CatBoost's performance improves for weekly predictions, with a strong R^2 of 0.744. However, monthly predictions show a sharp decline due to data inconsistencies, particularly during the imputation process. The model's performance drops sharply for monthly predictions, indicating a need for better data alignment. Figure 10 presents the regression plots for $PM_{2.5}$ across hourly, daily, weekly, and monthly time frames.

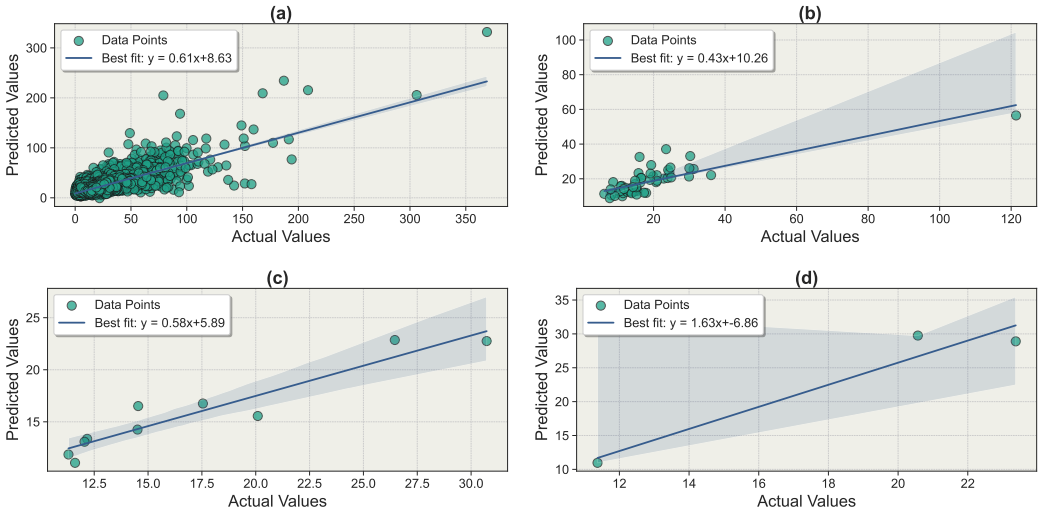

Figure 10. Regression plots of $PM_{2.5}$ pollutant for (**a**) hourly, (**b**) daily, (**c**) weekly, and (**d**) monthly time frames.

Finally, from the results shown in Table 7, CatBoost is a robust model for predicting PM_{10} concentrations across daily, weekly, and monthly time frames. It maintains consistent R^2 values above 0.7, making it well suited for modeling PM_{10} trends. The Extra Trees model shows strong performance for hourly predictions, while CatBoost's best results are observed for daily and weekly predictions, while monthly predictions remain strong. The regression plots for PM_{10} across hourly, daily, weekly, and monthly time frames are shown in Figure 11.

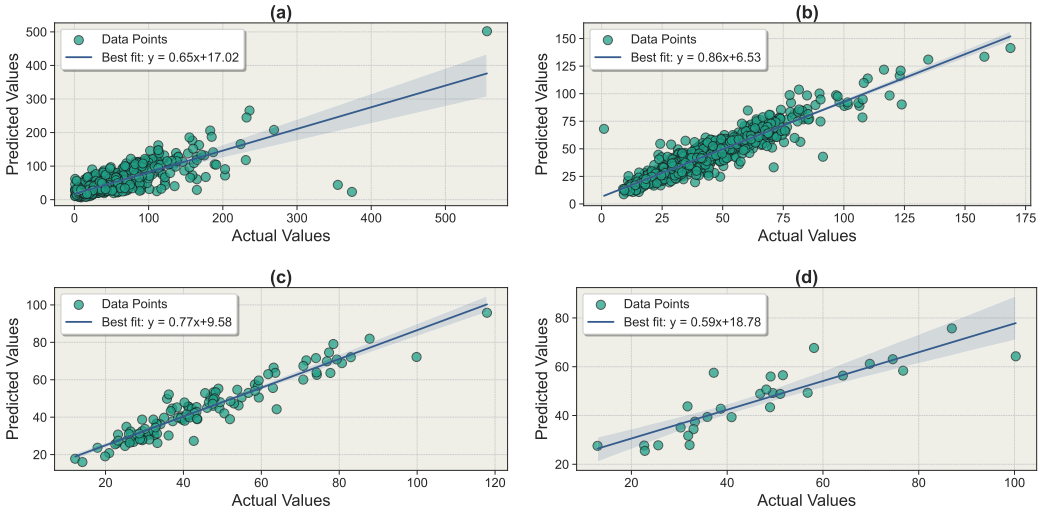

Figure 11. Regression plots of PM_{10} pollutant for (**a**) hourly, (**b**) daily, (**c**) weekly, and (**d**) monthly time frames.

4.2. ANOVA Analysis

In addition to obtaining the metrics of the different errors, both an ANOVA analysis and the Tukey test were performed. The ANOVA test allows us to discover whether the results obtained are significant, i.e., it allows us to determine whether it is necessary to reject the null hypothesis or accept the alternative hypothesis. Thus, in this section, the ANOVA analysis is presented for all the pollutants predicted.

The ANOVA test for the *CO* was conducted to assess whether differences in prediction performance across time frames were statistically significant. The results, presented in Table 8, reveal a significant *p*-value of 0.00171579, suggesting that the variation in model

performance between time frames is statistically significant. In addition, Figure 12a illustrates the error distribution for *CO* predictions across different time frames. The errors for hourly and daily predictions are relatively low, indicating strong model performance. However, the weekly and monthly time frames exhibit slightly higher errors, particularly for the weekly predictions.

Table 8. ANOVA analysis for criterion pollutants.

Pollutant	F-Statistic	*p*-Value
CO	5.06910789	0.00171579
O_3	4.71749567	0.00271908
SO_2	18.60966426	$4.69060098 \times 10^{-12}$
NO_2	2.5443607	0.0547634
$PM_{2.5}$	2.7779329	0.0396277
PM_{10}	38.5170041	$3.3244844 \times 10^{-24}$

For O_3, the results reveal a highly significant p-value of 0.00271908, indicating that the disparity in model performance over different time periods is statistically significant. Figure 12b shows that the errors for hourly and daily predictions are relatively low, indicating strong model performance. However, the weekly and monthly time frames exhibit slightly higher errors, particularly for the monthly predictions, but this is consistent with the lacking information for the monthly time frame

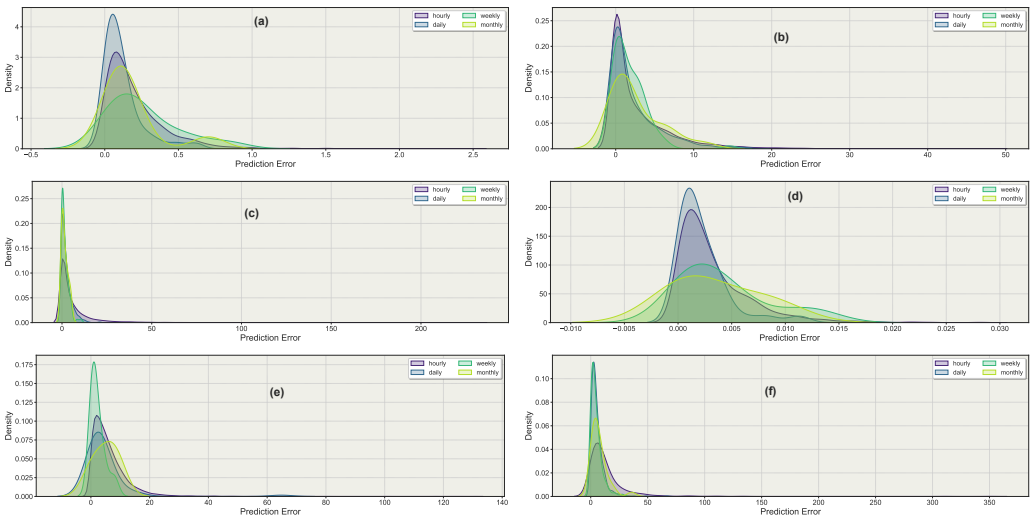

Figure 12. Error plot for the different time windows for (**a**) CO, (**b**) O_3, (**c**) SO_2, (**d**) NO_2, (**e**) $PM_{2.5}$, and (**f**) PM_{10}.

In the same Table 8, it is possible to observe that the results for SO_2 show a p-value of $4.69060098 \times 10^{-12}$, demonstrating that the variation in model performance across different time intervals is statistically significant. The broader error distribution in the hourly predictions further emphasizes the difficulty in modeling short-term variability for this pollutant. In contrast, the error distribution becomes narrower for weekly and monthly predictions, indicating more reliable model performance over longer periods. However, the monthly comparison cannot be reliable due to the inconsistencies in the data availability. Figure 12c shows the error distribution plot for SO_2 predictions across the different time windows.

The result for the ANOVA test that was conducted for NO_2 is presented in Table 8, and the obtained value is highly significant for a p-value of 0.0547634, demonstrating that the variation in model performance over time is statistically significant. In Figure 12d, it

is observed that the errors increase significantly for the weekly and monthly predictions, reflecting the models' inability to generalize well for longer-term NO_2 trends.

The ANOVA test that was performed for $PM_{2.5}$ show the error distribution for these predictions, as illustrated in Figure 12e. While the errors for hourly and daily predictions remain moderate, they increase significantly for monthly predictions, highlighting the model's challenges in capturing long-term $PM_{2.5}$ trends. The values obtained for this test are presented in Table 8.

Finally, the results obtained from the ANOVA test were analyzed to assess whether differences in prediction performance across time frames were statistically significant. The results are presented in Table 8 An F-statistic of 38.52 and a p-value of 3.32×10^{-24} also indicate a performance between time frames that is statistically significant. An illustration of the error distribution for PM_{10} predictions across different time frames is presented in Figure 12. The errors for daily and weekly predictions are relatively low, indicating strong model performance. However, the hourly and monthly time frames exhibit slightly higher errors, particularly for the monthly predictions, which is consistent with the observed R^2 values.

4.3. Tukey HSD

To identify which specific groups differ from each other, a post hoc Tukey's HSD test was performed. The results of this test are presented in Table 9. This study reveals a significant difference in the predictive performance between daily and hourly predictions for CO predictions, rejecting the null hypothesis that they exhibit the same performance. The models excelled in short-term forecasts, but the complexity of long-term predictions, exacerbated by data limitations, suggests the need for further refinement, possibly through alternative feature engineering techniques or enhanced imputation methods.

The results of the post hoc Tukey's HSD test suggest that while O_3 predictions for mid-term time frames are reasonably accurate, a noticeable decline in model performance is observed for the monthly time frame, indicating a difference in the accuracy of hourly predictions.

Also, the study conducted on SO_2 predictions found a significant difference in model performance between certain time frames, with a post hoc Tukey's HSD test confirming this difference. The results show that it is more challenging to perform predictions at shorter time frames due to its high variability, and the models show a stronger predictive ability for longer time frames, such as weekly and monthly, as the noise in the data is reduced.

In the case of the NO_2 prediction, Tukey's HSD test showed no significant differences between time frames, suggesting that despite some models' poor performance, the differences in prediction error are not significant enough to reject the null hypothesis. In addition, this study suggests that while models like CatBoost and Extra Trees perform well for short-term NO_2 predictions, significant improvements and adjustments are needed for successful modeling at longer time scales.

For the $PM_{2.5}$ predictions, with Tukey's HDS test, despite a drop in performance for monthly predictions, the differences were not significant enough to reject the null hypothesis. Here, CatBoost performs well for short-term and mid-term predictions of $PM_{2.5}$ but faces difficulties for monthly predictions due to imputation issues disrupting the detection of long-term trends.

Finally, the Tukey HSD test results reveal that CatBoost is the most effective model for predicting PM_{10} concentrations across daily, weekly, and monthly time frames, with robust R^2 values exceeding 0.7. This performance is statistically significant across time windows, confirming that PM_{10} trends can be reliably predicted, particularly for longer time frames.

Table 9. Tukey HSD test for criteria pollutants.

Pollutant	Group 1	Group 2	Mean Diference	p-adj	Lower	Upper	Reject
CO	Daily	Hourly	0.0876	0.0012	0.0271	0.148	True
	Daily	Monthly	0.0615	0.6582	−0.0761	0.1992	False
	Daily	Weekly	0.1427	0.0926	−0.0151	0.3006	False
	Hourly	Monthly	−0.026	0.951	−0.1516	0.0996	False
	Hourly	Weekly	0.0552	0.7709	−0.0923	0.2026	False
	Monthly	Weekly	0.0812	0.6985	−0.1112	0.2736	False
O_3	Daily	Hourly	0.7348	0.0091	0.1338	1.3357	True
	Daily	Monthly	0.4449	0.9746	−2.2578	3.1476	False
	Daily	Weekly	−0.3433	0.9264	−1.7708	1.0843	False
	Hourly	Monthly	−0.2898	0.9922	−2.931	2.3513	False
	Hourly	Weekly	−1.078	0.1471	−2.3853	0.2292	False
	Monthly	Weekly	−0.7882	0.9015	−3.7297	2.1533	False
SO_2	Daily	Hourly	3.819	0.0	2.2994	5.3387	True
	Daily	Monthly	−0.7812	0.9924	−7.9753	6.4128	False
	Daily	Weekly	−0.7979	0.9487	−4.5811	2.9853	False
	Hourly	Monthly	−4.6003	0.3344	−11.6365	2.436	False
	Hourly	Weekly	−4.617	0.0036	−8.0907	−1.1433	True
	Monthly	Weekly	−0.0167	1.0	−7.8597	7.8263	False
NO_2	Daily	Hourly	0.0011	0.0515	−0.0	0.0021	False
	Daily	Monthly	0.0013	0.9205	−0.0038	0.0063	False
	Daily	Weekly	0.002	0.2297	−0.0007	0.0047	False
	Hourly	Monthly	0.0002	0.9996	−0.0048	0.0052	False
	Hourly	Weekly	0.0009	0.7727	−0.0016	0.0034	False
	Monthly	Weekly	0.0007	0.9868	−0.0048	0.0063	False
$PM_{2.5}$	Daily	Hourly	1.7944	0.1486	−0.3865	3.9754	False
	Daily	Monthly	0.5548	0.999	−9.4209	10.5305	False
	Daily	Weekly	−2.2374	0.7505	−7.9969	3.5221	False
	Hourly	Monthly	−1.2396	0.9879	−10.9758	8.4966	False
	Hourly	Weekly	−4.0319	0.2105	−9.3657	1.302	False
	Monthly	Weekly	−2.7922	0.9169	−13.8922	8.3077	False
PM_{10}	Daily	Hourly	7.4499	0.0	5.5461	9.3536	True
	Daily	Monthly	1.8781	0.9209	−5.7336	9.4898	False
	Daily	Weekly	−0.9476	0.9316	−4.9989	3.1038	False
	Hourly	Monthly	−5.5717	0.2299	−13.1262	1.9827	False
	Hourly	Weekly	−8.3974	0.0	−12.3402	−4.4547	True
	Monthly	Weekly	−2.8257	0.8208	−11.1838	5.5325	False

5. Discussion

The findings of this study underscore the significance of time frame selection in predicting air pollution in Salamanca, Mexico. It was observed that prediction models exhibited superior performance on longer time scales, such as weekly and monthly predictions, in comparison to hourly and daily predictions. This observation suggests that short-term variations in pollutant concentrations may introduce greater variability in the models, potentially compromising their accuracy and reliability.

Additionally, it was noted that certain pollutants pose significant modeling challenges. For instance, ozone (O_3) and sulfur dioxide (SO_2) exhibited relatively high accuracy across diverse time frames. In contrast, models for nitrogen dioxide (NO_2) and particulate matter ($PM_{2.5}$ and PM_{10}) demonstrated greater variability in their performance. This variability may be attributed to emission sources and their interaction with meteorological variables

such as humidity and temperature, underscoring the necessity for more robust modeling approaches that incorporate additional contextual information.

The employment of low-code AutoML tools, such as PyCaret, facilitated the model selection and comparison process, thereby enabling the identification of the most suitable algorithms for each context. In this regard, the performance of the Extra Trees regressor and CatBoost stood out, both of which exhibited superior accuracy and stability in comparison to other models. However, the findings also underscore the significance of data preprocessing and missing value imputation, as inconsistencies were identified that constrained performance in specific scenarios.

Conversely, correlation analysis between meteorological variables and pollutant concentrations underscored the influence of factors such as wind speed and solar radiation on pollutant dispersion. These findings underscore the necessity to incorporate multiple environmental variables into predictive models and to investigate strategies such as hybrid models or deep learning techniques to enhance prediction accuracy over shorter time frames.

This study has also demonstrated the potential of combining temporal, meteorological and pollutant data in air quality prediction. Similarly, it has shown the value of low-code AutoML tools in facilitating model development and improving prediction accuracy. However, significant challenges remain, particularly with regard to handling data inconsistencies and expanding modeling methodologies.

In addition, data quality plays a critical role in the accuracy of air pollution prediction models. In this study, several inconsistencies in the data used were identified, which pose challenges to the stability and reliability of the predictions. These inconsistencies are mainly missing values; however, noise in the measurements, variability in temporal resolution, and the possible omission of key environmental factors can equally affect the results.

Therefore, the main problem detected is the presence of missing values, which affects the ability of the models to identify consistent patterns in the data. Missing information on meteorological variables and pollutant concentrations can generate biases in the results and reduce the accuracy of machine learning algorithms. Thus, data imputation techniques were implemented, but it is clear that more advanced methods could improve the stability of the predictions, which will be considered for further investigations.

As a kind of conclusion, we must comment that ozone (O_3) is the most predictable pollutant since its performance was fairly consistent across time scales, making it the easiest to forecast. The reason is that ozone is highly dependent on solar energy and temperature, two meteorological variables that show more stable trends with time. Meanwhile, the most complex pollutants to model were particulate matter ($PM_{2.5}$ and PM_{10}) and nitrogen dioxide (NO_2), which were the hardest pollutants to predict. This is because their concentration is affected by numerous immeasurable factors, including emissions from vehicles, industrial processes, and sudden shifts in the weather. Furthermore, these contaminants' variations are more pronounced in shorter periods, which adds noise to the models and diminishes their precision.

Future studies will be oriented toward developing cascade systems or combinations of models to improve prediction accuracy by leveraging the strengths of different algorithms. In addition, it is essential to address the issue of missing data by implementing real-time mechanisms to detect and correct these inconsistencies, either through advanced imputation techniques or predictive models capable of efficiently completing the information.

Finally, extending this research to other geographical locations, especially in urban areas with different environmental conditions, would allow the robustness of the models to be validated and universal patterns or specific behaviors of each pollutant to be determined.

These advances would contribute to the improvement of air quality management strategies and the formulation of evidence-based public policies.

Author Contributions: Investigation, writing—original draft preparation, software, and validation, F.-J.M.-V.; conceptualization, project administration, writing—review and editing, and supervision, F.T.-R.; investigation, methodology, and writing—original draft preparation, A.E.V.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Data are contained within the article. Further inquiries can be directed to the corresponding author(s).

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

References

1. Jiang, P.; Dong, Q.; Li, P. A novel hybrid strategy for $PM_{2.5}$ concentration analysis and prediction. *J. Environ. Manag.* **2017**, *196*, 443–457. [CrossRef]
2. Sinaica, S.; INECC. Principios de Medición de la Calidad del Aire. 2018 Available online: http://sinaica.inecc.gob.mx/pags/guias.php (accessed on 19 September 2024).
3. Ubilla, C.; Yohannessen, K. Contaminación Atmosférica Efectos en la Salud Respiratoria en el Niño. *Rev. Médica Clínica Las Condes* **2017**, *28*, 111–118. [CrossRef]
4. Etchie, T.O.; Etchie, A.T.; Adewuyi, G.O.; Pillarisetti, A.; Sivanesan, S.; Krishnamurthi, K.; Arora, N.K. The gains in life expectancy by ambient $PM_{2.5}$ pollution reductions in localities in Nigeria. *Environ. Pollut.* **2018**, *236*, 146–157. [CrossRef] [PubMed]
5. Masse, F. La Contaminación del Aire: Un Problema que daña la Salud y la Economía. 2022. Available online: https://imco.org.mx/la-contaminacion-del-aire-un-problema-que-dana-la-salud-y-la-economia/ (accessed on 13 August 2024).
6. Talbi, A.; Kerchich, Y.; Kerbachi, R.; Boughedaoui, M. Assessment of annual air pollution levels with PM_1, $PM_{2.5}$, PM_{10} and associated heavy metals in Algiers, Algeria. *Environ. Pollut.* **2018**, *232*, 252–263. [CrossRef] [PubMed]
7. Son, Y.; Osornio-Vargas, A.R.; O'Neill, M.S.; Hystad, P.; Texcalac-Sangrador, J.L.; Ohman-Strickland, P.; Meng, Q.; Schwander, S. Land use regression models to assess air pollution exposure in Mexico City using finer spatial and temporal input parameters. *Sci. Total Environ.* **2018**, *639*, 40–48. [CrossRef]
8. Orioli, R.; Cremona, G.; Ciancarella, L.; Solimini, A.G. Association between PM_{10}, $PM_{2.5}$, NO_2, O_3 and self-reported diabetes in Italy: A cross-sectional, ecological study. *PLoS ONE* **2018**, *13*, e0191112. [CrossRef] [PubMed]
9. WHO. New WHO Global Air Quality Guidelines Aim to Save Millions of Lives from Air Pollution. 2021. Available online: https://www.who.int/news/item/22-09-2021-new-who-global-air-quality-guidelines-aim-to-save-millions-of-lives-from-air-pollution (accessed on 21 September 2024).
10. Narkhede, G.; Deore, A.; Kolte, B. Machine Learning Models for daily AQI Prediction: An In-depth Analysis. In Proceedings of the 2024 First International Conference on Innovations in Communications, Electrical and Computer Engineering (ICICEC), Davangere, India, 24–25 October 2024; pp. 1–6. [CrossRef]
11. Uddin, R.; Faiaz, A.; Islam, S.R. Forecasting Air Quality by Estimating $PM_{2.5}$ Concentration Level Using k-Nearest Neighbor Model in Gazipur, Bangladesh. In *Advances in Civil Engineering Materials*; Nia, E.M., Awang, M., Eds.; Springer Nature: Singapore, 2024; pp. 189–198. [CrossRef]
12. Borah, J.; Kumar, S.; Kumar, N.; Nadzir, M.S.M.; Cayetano, M.G.; Ghayvat, H.; Majumdar, S.; Kumar, N. AiCareBreath: IoT-Enabled Location-Invariant Novel Unified Model for Predicting Air Pollutants to Avoid Related Respiratory Disease. *IEEE Internet Things J.* **2024**, *11*, 14625–14633. [CrossRef]
13. Narkhede, G.; Hiwale, A. Deep learning approach to forecast air pollution based on novel hourly index. *Phys. Scr.* **2023**, *98*, 096004. [CrossRef]
14. Arnaut, F.; Đurđević, V.; Kolarski, A.; Srećković, V.A.; Jevremović, S. Improving Air Quality Data Reliability through Bi-Directional Univariate Imputation with the Random Forest Algorithm. *Sustainability* **2024**, *16*, 7629. [CrossRef]
15. Zareba, M.; Cogiel, S.; Danek, T.; Weglinska, E. Machine Learning Techniques for Spatio-Temporal Air Pollution Prediction to Drive Sustainable Urban Development in the Era of Energy and Data Transformation. *Energies* **2024**, *17*, 2738. [CrossRef]
16. Ksibi, A.; Salhi, A.; Alluhaidan, A.S.; El-Rahman, S.A. Airstacknet: A Stacking Ensemble-Based Approach for Air Quality Prediction. *Comput. Mater. Contin.* **2022**, *74*, 2073–2096. [CrossRef]
17. Rad, A.K.; Nematollahi, M.J.; Pak, A.; Mahmoudi, M. Predictive modeling of air quality in the Tehran megacity via deep learning techniques. *Sci. Rep.* **2025**, *15*, 1367. [CrossRef] [PubMed]

18. Westergaard, G.; Erden, U.; Mateo, O.A.; Lampo, S.M.; Akinci, T.C.; Topsakal, O. Time Series Forecasting Utilizing Automated Machine Learning (AutoML): A Comparative Analysis Study on Diverse Datasets. *Information* **2024**, *15*, 39. [CrossRef]
19. Mazahir Mehdi Zaidi, S.; Rizvi, S.; Arora, D.; Tiwari, S. An AutoML approach integrated with live weather data in rain forecasting system (RFS). In *Inventive Communication and Computational Technologie*; Ranganathan, G., Papakostas, G.A., Shi, Y., Eds.; Springer Nature: Singapore, 2024; pp. 43–52. [CrossRef]
20. Yu, M.; Zhang, S.; Zhang, K.; Yin, J.; Varela, M.; Miao, J. Developing high-resolution $PM_{2.5}$ exposure models by integrating low-cost sensors, automated machine learning, and big human mobility data. *Front. Environ. Sci.* **2023**, *11*, 1223160. [CrossRef]
21. Quispe, J.O.Q.; Quispe, A.C.F.; Calvo, N.C.L.; Toledo, O.C. Analysis and Selection of Multiple Machine Learning Methodologies in PyCaret for Monthly Electricity Consumption Demand Forecasting. *Mater. Proc.* **2024**, *18*, 5. [CrossRef]
22. Molepo, K.M.; Abiodun, B.J.; Magoba, R.N. The transport of PM_{10} over Cape Town during high pollution episodes. *Atmos. Environ.* **2019**, *213*, 116–132. [CrossRef]
23. INEGI. México en Cifras. 2021. Available online: https://www.inegi.org.mx/app/areasgeograficas/ (accessed on 20 August 2024).
24. Alimissis, A.; Philippopoulos, K.; Tzanis, C.G.; Deligiorgi, D. Spatial estimation of urban air pollution with the use of artificial neural network models. *Atmos. Environ.* **2018**, *191*, 205–213. [CrossRef]
25. Wilson, F.R. Mexican Refinery Lines Impoundment with Geomembrane. 2007. Available online: https://web.archive.org/web/20070816005106/http://www.seamannewsroom.com/newsreleases/pemex_article.html (accessed on 5 July 2024).
26. Greimel, H.; News, A. Mazda Opens New Plant in Salamanca, Mexico. 2014. Available online: https://www.autoweek.com/news/a1888996/mazda-opens-new-plant-salamanca-mexico/ (accessed on 5 July 2024).
27. de Medio Ambiente y Ordenamiento Territorial, S. Guanajuato | Secretaría del Medio Ambiente y Ordenamiento Territorial. Available online: https://smaot.guanajuato.gob.mx/sitio/historicos/ (accessed on 5 July 2024).
28. Pan, K.; Lu, J.; Li, J.; Xu, Z. A Hybrid Autoformer Network for Air Pollution Forecasting Based on External Factor Optimization. *Atmosphere* **2023**, *14*, 869. [CrossRef]
29. Ferrari, L.; Guariso, G. Geography-based Neural Networks for the Simulation of Air Pollution. In Proceedings of the IFAC-PapersOnLine, Canberra, Australia, 4–6 January 2023; Volume 56. [CrossRef]
30. Haddad, K.; Vizakos, N. Air quality pollutants and their relationship with meteorological variables in four suburbs of Greater Sydney, Australia. *Air Qual. Atmos. Health* **2021**, *14*, 55–67. [CrossRef]
31. Cai, J.; Luo, J.; Wang, S.; Yang, S. Feature selection in machine learning: A new perspective. *Neurocomputing* **2018**, *300*, 70–79. [CrossRef]
32. Ali, M. *PyCaret: An Open Source, Low-Code Machine Learning Library in Python*, PyCaret Version 3.0. 2020 . Available online: https://pycaret.org/ (accessed on 21 September 2024).
33. Akiba, T.; Sano, S.; Yanase, T.; Ohta, T.; Koyama, M. Optuna: A Next-generation Hyperparameter Optimization Framework. In Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Anchorage, AK, USA, 4–8 August 2019. [CrossRef]
34. Bergstra, J.; Komer, B.; Eliasmith, C.; Yamins, D.; Cox, D.D. Hyperopt: A Python library for model selection and hyperparameter optimization. *Comput. Sci. Discov.* **2015**, *8*, 014008. [CrossRef]
35. Zahedi, L.; Mohammadi, F.G.; Rezapour, S.; Ohland, M.W.; Amini, M.H. Search Algorithms for Automated Hyper-Parameter Tuning. *arXiv* **2021**, arXiv:2104.14677.
36. Bhat, P.C.; Prosper, H.B.; Sekmen, S.; Stewart, C. Optimizing event selection with the random grid search. *Comput. Phys. Commun.* **2018**, *228*, 245–257. [CrossRef]
37. PyCaret. Quickstart | Docs. Available online: https://pycaret.gitbook.io/docs/get-started/quickstart (accessed on 12 September 2024).

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.