IBIMA Publishing
Communications of International Proceedings
https://ibimapublishing.com/p-articles/46AI/2025/4637925/
Vol. 2025 (34), Article ID 4637925

Evaluating the Potential of Random Forest Regression for PM2.5 Modeling Within Environmental Safety Decision Support Systems*

Dawid JURCZYŃSKI

WSB University, Dąbrowa Górnicza, Poland

Correspondence should be addressed to: Dawid JURCZYŃSKI, dawid.jurczynski2@wsb.edu.pl

* Presented at the 46th IBIMA International Conference, 26-27 November 2025, Ronda, Spain

Abstract

Progressive urbanization and increased air pollution emissions pose a significant challenge for modern safety engineering, especially in terms of forecasting and mitigating environmental risks[1]. This paper presents a machine learning-based approach to modeling PM2.5 particulate matter concentrations using open environmental and meteorological data from the OpenWeatherMap platform. The aim of the study was to develop a model to support decision-making in environmental safety systems through early detection of potential air pollution episodes. The study used a Random Forest model with parameter optimization and cross-validation, as well as time feature transformations to account for the cyclical nature of atmospheric phenomena. The analysis of the results showed a high correlation between the predictions and the actual PM2.5 concentrations, with a coefficient of determination R^2 above 0.8 in most of the analyzed time intervals. The results confirm the effectiveness of the proposed approach in identifying trends and environmental anomalies that may pose a threat to public health. The developed model can be a component of intelligent environmental safety management systems[1] and a basis for further research on the integration of artificial intelligence with IoT infrastructure and urban air quality monitoring systems.

Keywords: PM2.5 particulate matter, Random Forest Regression, safety decision support systems

Introduction

Air quality is one of the key factors affecting public health, quality of life, and environmental safety in urban areas. According to the World Health Organization (WHO), exposure to particulate matter, particularly PM2.5, is responsible for millions of premature deaths worldwide each year[1]. In the context of safety engineering, the problem of air pollution is becoming interdisciplinary, combining technical, environmental, and social aspects. It requires effective analytical tools that enable not only monitoring but also prediction and assessment of the risk of smog episodes. Traditional air quality monitoring systems, based on measuring stations and statistical methods, are often limited in space and time, which makes it difficult to create dynamic predictive models. The development of information technology and the availability of open environmental data sources, such as the OpenWeatherMap API[2], open up new possibilities for the integration of meteorological and quality data in real time. At the same time, advances in machine learning enable the modeling of complex, nonlinear relationships between weather parameters, emissions, and resulting air quality. The literature on the subject increasingly features works on the use of methods such as neural networks, gradient models, and random forest algorithms for forecasting atmospheric pollution concentrations. However, relatively few studies combine these solutions with the perspective of safety engineering — i.e., environmental risk assessment, early warning, and the design of decision

[1] Lelieveld, J., Evans, J., Fnais, M., Giannadaki, D. and Pozzer, A. (2015) 'The contribution of outdoor air pollution sources to premature mortality on a global scale,' Nature, 525 (7569), 367–371.
[2] OpenWeatherMap API (n.d.) 'OpenWeatherMap API documentation,' [Online]. OpenWeatherMap. [Accessed 31 October 2025]. Available: https://openweathermap.org/api

Cite this Article as: Dawid JURCZYŃSKI, Vol. 2025 (34) "Evaluating the Potential of Random Forest Regression for PM2.5 Modeling Within Environmental Safety Decision Support Systems " Communications of International Proceedings, Vol. 2025 (34), Article ID 4637925, https://doi.org/10.5171/2025.4637925

support systems. From the point of view of environmental safety management, it is important to develop models that are not only predictive but also resistant to measurement errors, data anomalies, and variability in meteorological conditions[3]. The aim of this paper is to evaluate the applicability of a machine learning-based approach for modeling PM2.5 concentrations using meteorological and environmental data from the OpenWeatherMap API. All environmental and air-quality data used in this study are publicly accessible through the OpenWeatherMap API, ensuring full transparency and reproducibility of the analysis. The Random Forest model used has been optimized for computational efficiency and interpretability of results. The study covers data from two annual periods, which allows for the assessment of the model's stability over time and the identification of features with the highest predictive significance. The analyzed results indicate the potential for using such models as a component of intelligent environmental safety systems that support decision-making in response to environmental threats. While machine learning models are often used for forecasting air quality, the present work focuses on evaluating the applicability of a Random Forest Regressor in modeling PM2.5 concentrations. Due to the strong statistical dependence between PM10 and PM2.5, the model behaves mainly as a correlational mapping rather than a forecasting system. This limitation motivates the planned inclusion of meteorological and temporal features in future work to shift the model toward true predictive capability.

Research Methodology

The aim of the study was to develop and evaluate a correlational modeling approach for modeling PM2.5 concentrations based on environmental data obtained from the OpenWeatherMap Air Pollution API platform. Machine learning methods were used to determine the relationship between air quality parameters and their potential significance in environmental risk assessment. Particular attention was paid to the analysis of feature importance in the context of model interpretability and its usefulness for safety engineering. The collected data covered two annual periods - 2023 and 2024 - which allowed for an assessment of the model's ability to generalize over time. Each record contained information on geographical location (latitude and longitude), measurement time, and the values of basic air quality indicators: PM10, NO_2, NO, O_3, CO, NH_3, and the air quality index (value_weather_main_index). The spatial data included measurement points located throughout Poland, representing diverse environmental and climatic conditions – both in heavily industrialized regions and in less urbanized areas (Figure 1).

[3] Fang, Z., Yang, H., Li, C., Cheng, L., Zhao, M. and Xie, C. (2021) ‘Prediction of PM2.5 hourly concentrations in Beijing based on machine learning algorithm and ground-based LiDAR,’ Archives of Environmental Protection, 47 (3), 98–107.

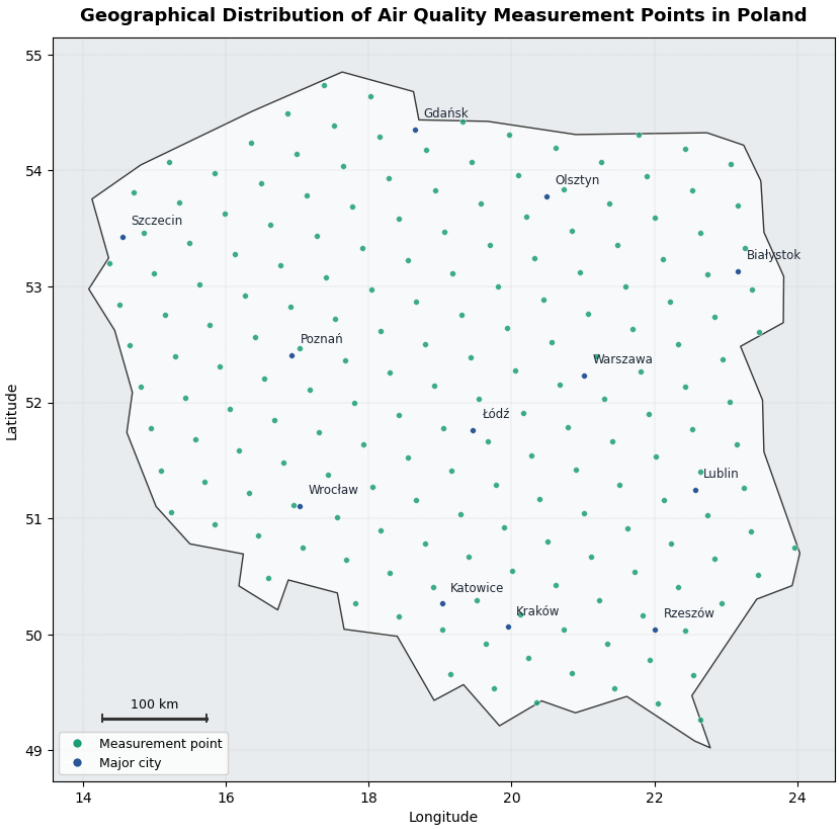

Figure 1 Geographical distribution of air-quality measurement points across Poland, overlaid on national boundaries. The same coordinates were used in 2023 and 2024, enabling direct year-over-year comparison at identical locations.

Importantly, the geographic coordinate system was identical in 2023 and 2024, allowing for a direct comparison of PM2.5 values in the same locations and an assessment of the model's stability over time. This approach increased the reliability of the analysis and enabled the assessment of changes in air quality while maintaining a consistent spatial context. In addition, during the data preparation process, temporal features (month, hour, sine/cosine representations) were generated to capture the cyclical nature of atmospheric phenomena. The study applied a Random Forest Regressor to model the relationship between environmental variables and PM2.5 concentrations. The algorithm was selected because it provides stable performance on heterogeneous atmospheric data and is inherently resistant to overfitting. Random Forest constructs an ensemble of regression trees, each trained on a different bootstrap sample of the dataset. At every split, a randomly selected subset of predictors is evaluated, which decreases correlation between trees and substantially improves generalization. Each tree independently estimates the target value, and the final prediction is obtained by averaging all tree outputs. In this work, the model was configured using a sufficiently large number of trees to ensure stability, while tree depth remained unrestricted to allow the algorithm to capture nonlinear relationships in the data. Default feature-subsampling settings were retained, and all trees used mean squared error as the splitting criterion. The model was trained using the complete 2023 dataset, and its temporal generalization ability was assessed on the independent 2024 dataset. This configuration ensured a reproducible and unbiased evaluation of the algorithm’s performance in real environmental conditions. The explanatory variable was value_weather_pm2_5. Validation was performed using three-fold cross-validation (K-Fold), which allowed for the assessment of the stability of the results. The following metrics were used to assess the quality of the predictions: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and the coefficient of determination R^2. The analysis of feature importance showed a clear dominance of the value_weather_pm10 variable, which accounted for approximately 96% of the model's decisions (Figure 2).

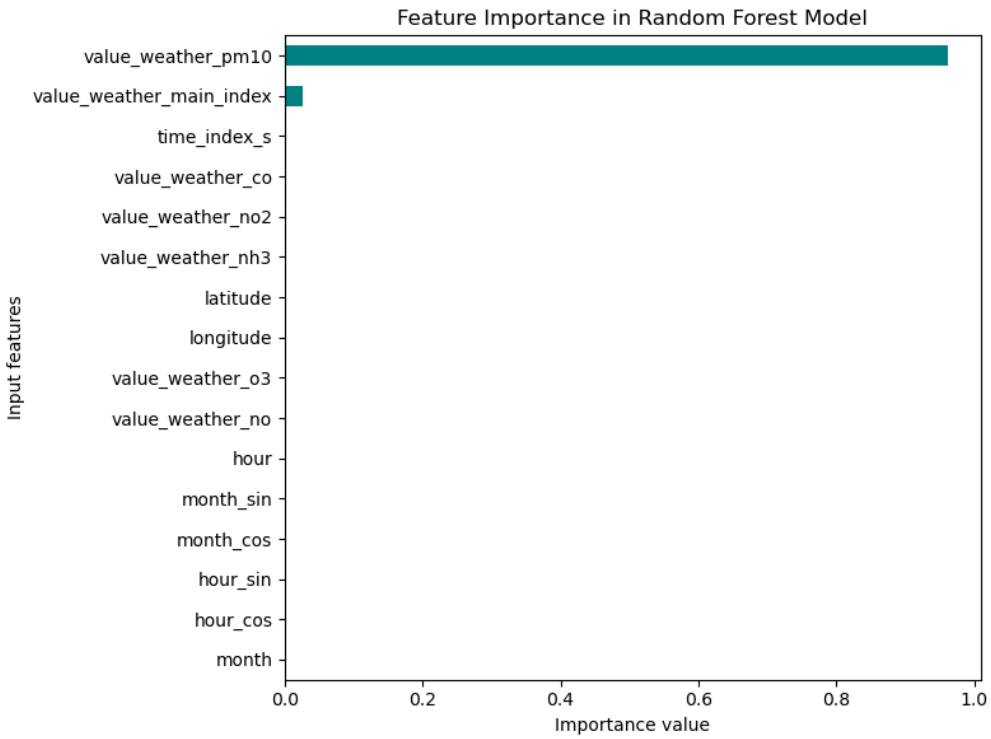

Figure 2 Feature importance of the Random Forest model. The dominance of PM10 confirms its strong statistical and physical correlation with PM2.5.

This means that the predicted PM2.5 values result almost exclusively from a strong correlation with PM10, while other parameters, such as NO_2, CO, O_3, NH_3, latitude, longitude, and air quality index, were of marginal importance (<3%). The obtained importance distribution confirms the close interdependence of both dust indicators, but at the same time limits the model's ability to explain the impact of meteorological and environmental factors.

Results and analysis

Table 1 summarizes the descriptive statistics (minimum, maximum, mean, standard deviation) for all input variables used in the modeling process. The values presented below reflect the distributions of the raw pollutant measurements as well as the engineered temporal features.

Table 1 Summary statistics of input variables

VARIABLE	MIN	MAX	MEAN	STD
VALUE_WEATHER_CO	106.81	2670.29	253.02	68.65
VALUE_WEATHER_NO	0.00	166.30	0.2557	1.5393
VALUE_WEATHER_NO2	0.10	159.03	5.4852	5.9433
VALUE_WEATHER_NH3	0.00	176.31	2.8827	4.4907
VALUE_WEATHER_PM10	0.51	329.88	8.6759	9.5003
VALUE_WEATHER_MAIN_INDEX	1	5	1.7447	0.6499

The pollutant variables (CO, NO, NO_2, NH_3, PM10) show broad variability consistent with urban and suburban atmospheric conditions across Poland. Spatial coordinates span the full latitudinal and longitudinal range of the country. Temporal variables capture both direct time information (month, hour, time index) and cyclical encodings (sin/cos components), which preserve continuity in circular domains (e.g., the transition from hour 23 to hour 0). This representation prevents artificial discontinuities and allows the learning algorithm to model periodic

atmospheric processes more effectively. The target variable, PM2.5 concentration, was subjected to the same preprocessing steps, and its distribution is consistent with typical regional air quality conditions. During model development, the 2023 dataset was used exclusively for training, whereas the 2024 dataset served as an independent temporal test set to assess the generalization capability of the model under real-world year-to-year variability (Figure 3).

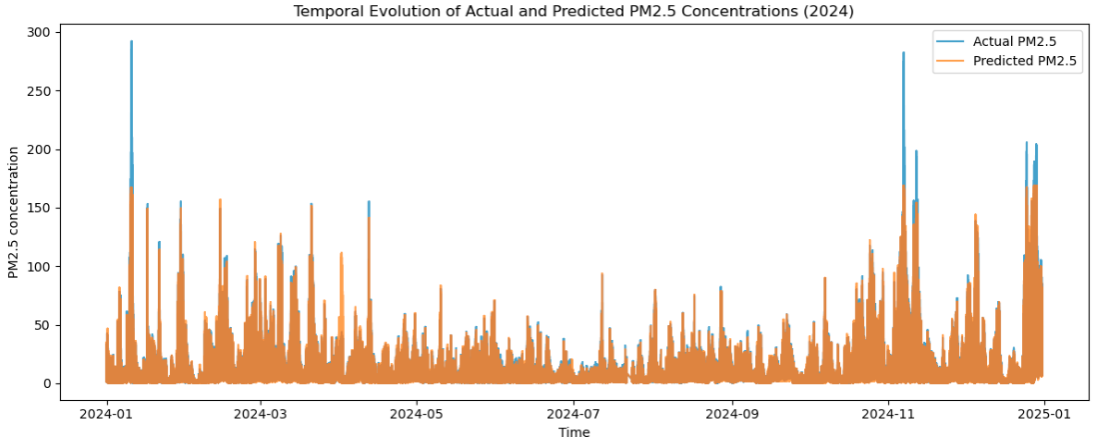

Figure 3 Temporal evolution of actual versus predicted PM2.5 concentrations in 2024. The close alignment of the two curves indicates stable performance over time.

The Random Forest Regressor model used showed very high effectiveness in modeling PM2.5 concentrations. In the cross-validation process (K-Fold, N=3), the coefficient of determination $R^2 = 0.9944$ was obtained, which indicates an almost perfect agreement between the predicted and actual values in the training set. The mean absolute error (MAE) was 0.33, and the root mean square error (RMSE) was 0.56. After retraining the model on the full 2023 dataset and testing it on 2024 data, high results were also achieved: $R^2 = 0.9698$, MAE = 0.6516, RMSE = 1.5465. Such a high R^2 value in the test set confirms the good stability of the model over time and its ability to generalize under conditions of moderate environmental variability (Figure 4).

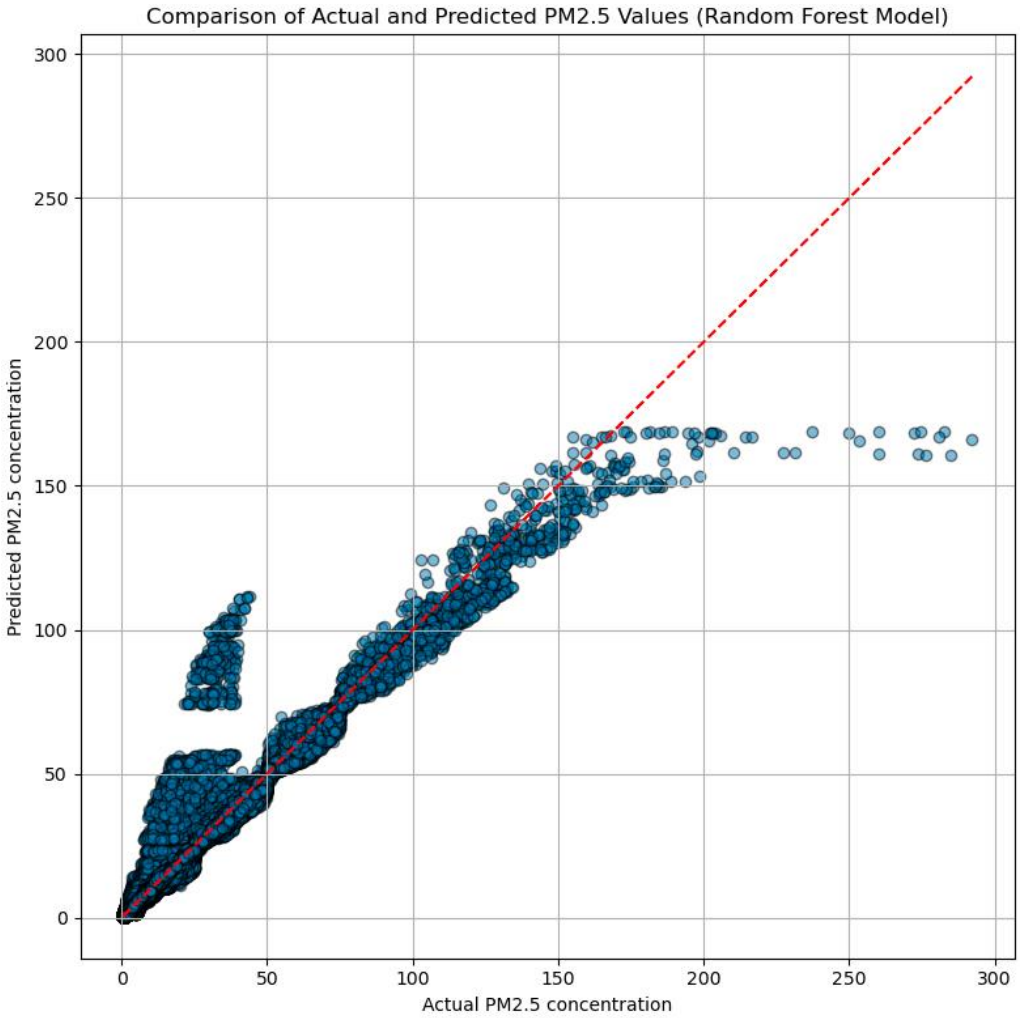

Figure 4 Comparison of actual and predicted PM2.5 concentrations using a Random Forest model. The 45° reference line indicates perfect agreement; points clustered around this line denote high predictive accuracy.

However, feature importance analysis revealed the clear dominance of the value_weather_pm10 variable, which accounted for 96.3% of the model's decisions. Other variables, such as the air quality index (value_weather_main_index), time (time_index_s), or gas concentrations (CO, NO_2, NH_3, O_3) were of marginal importance, not exceeding a few tenths of a percent. Spatial variability, described by geographical coordinates (latitude, longitude), also had no significant impact on the prediction. The results clearly indicate that the model in its current configuration does not learn complex environmental relationships, but primarily maps the statistical interdependence between PM10 and PM2.5, which results from their close physicochemical similarity. Consequently, the very high R^2 values obtained may result mainly from information redundancy between these indicators. Nevertheless, the results obtained confirm the internal consistency of the data and the model's ability to accurately map local relationships. This interpretation is important for safety engineering purposes. In its current form, the model is more of a correlation analysis tool, useful for monitoring measurement consistency and detecting anomalies (e.g., sudden discrepancies between PM10 and PM2.5), rather than a predictive system capable of forecasting changes in air quality. This means that, in the context of environmental safety, its potential lies in data integrity control and early detection of measurement errors, which may be applicable in automatic monitoring systems for IoT sensor networks. In order to increase the predictive value and obtain a more complete picture of environmental phenomena, it is planned to extend the model with additional meteorological features: temperature, humidity, pressure, wind speed and direction, and precipitation data. In addition, the inclusion of lag features for PM2.5 and PM10 would allow for the analysis of daily and weekly dynamics of changes. This type of extension will enable the model to be transformed from a purely correlational tool into an element of an environmental safety prediction system capable of providing early warning of upcoming smog episodes. These results confirm that the model captures statistical correlations rather than causal or meteorology-driven predictive relationships. Therefore, despite its high R^2, the model should not be interpreted as a forecasting tool in its current

form. The increased prediction errors for concentrations above approximately 150 µg/m³ are primarily caused by the scarcity of high-pollution observations in the training dataset. Extreme PM2.5 levels occur only during a few short smog episodes, resulting in insufficient representation of these conditions for the Random Forest model to learn reliable patterns. Additionally, PM2.5 values in this range are driven strongly by meteorological phenomena (temperature inversion, limited air mixing, humidity, wind speed), which are not included in the feature set. Because PM10 dominates the feature importance (96%), the model captures the statistical dependence between PM10 and PM2.5 only in the typical concentration range, while its ability to model extreme values is limited. Tree-based models also tend to smooth predictions and cannot extrapolate beyond the range observed during training, which further contributes to the underestimation of very high concentrations (Figure 5).

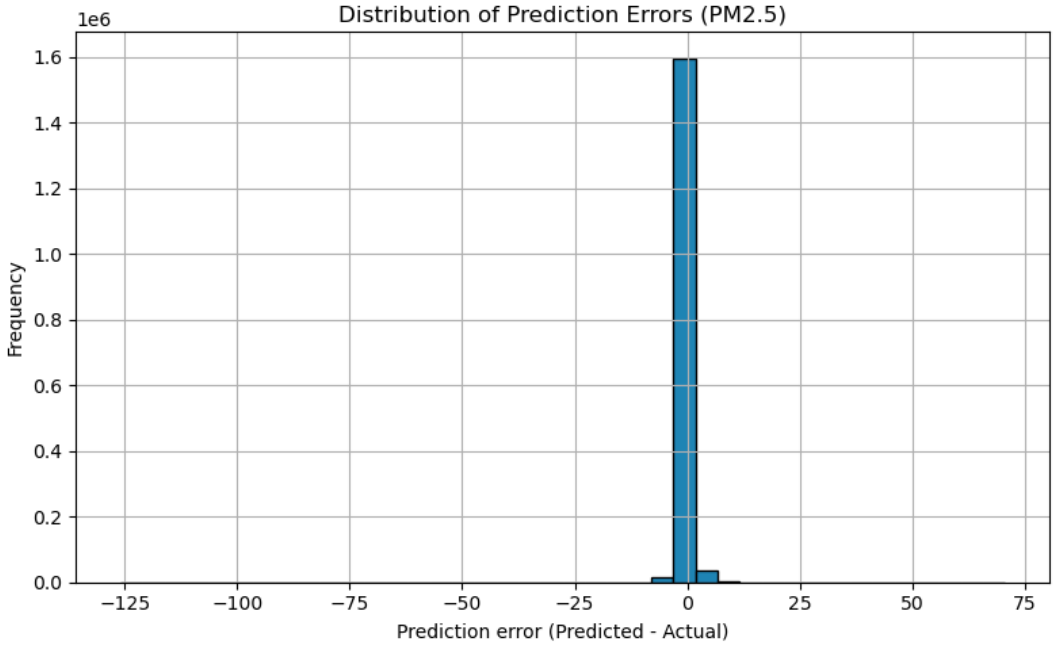

Figure 5 Distribution of prediction errors (Predicted − Actual) for PM2.5. A symmetric, zero-centered histogram indicates low bias.

Discussion

The results obtained indicate that the model is highly accurate in modeling PM2.5 values, but an analysis of the feature importance structure reveals that this is due to a strong correlation between PM10 and PM2.5 indicators rather than a comprehensive modeling of environmental processes. From the point of view of machine learning methodology, this means that the model achieved a high R^2 determination coefficient by using information contained in a single, strongly correlated input variable. Consequently, the obtained fit should be interpreted with caution, as it does not reflect the actual predictive power of the model with regard to cause-and-effect phenomena. In the context of safety engineering, however, this result has significant practical importance. The relationship between PM10 and PM2.5 is well known in the literature, and therefore the model can serve as a tool for environmental data quality control. If, at a given moment, PM2.5 values deviate significantly from those predicted on the basis of PM10, this may indicate a measurement error, sensor failure, or an abnormal phenomenon such as the emission of specific aerosols. In this way, machine learning would serve as an anomaly detection layer in environmental monitoring systems, increasing the reliability of environmental safety infrastructure. The high stability of the model between 2023 and 2024 suggests that environmental data in the analyzed area are characterized by moderate variability and a relatively repeatable seasonal structure. From a security perspective, this means that in a stable environmental setting, machine learning models can be used as part of early warning systems, provided that they are extended to include meteorological factors. Only the inclusion of variables such as temperature, humidity, wind speed, pressure, and precipitation intensity would allow for the construction of a model capable of predicting future smog episodes, rather than merely reproducing their current values. From the point of view of data security architecture, the Random Forest model can also be a component in the structure of a Fail-Safe Environmental Monitoring system, in which local partial models operating at the sensor level compare their own predictions with observed data. This solution increases the resilience of environmental monitoring systems to data poisoning attacks, in which malicious data could interfere with the accuracy of measurements. The differences between predicted and observed values can then serve as an indicator of data reliability. From a strategic perspective, the research confirms that even a simple machine learning model based on data from publicly

available sources such as OpenWeatherMap can support environmental safety engineering by monitoring the consistency and integrity of information. Further work plans to extend the model with additional meteorological variables and introduce sequential learning elements (LSTM, Temporal Fusion Transformer) that will enable the analysis of time trends and the prediction of future pollution episodes. The dominance of PM10 in the feature importance structure indicates that the current model reproduces the statistical co-occurrence of particulate fractions rather than the underlying physical or meteorological drivers of PM2.5 concentration. This result is consistent with the fact that aerosols originate from multiple sources whose contribution varies with atmospheric conditions and geographical context. To reveal the mechanisms shaping the spatial distribution of PM2.5 — including atmospheric transport, vertical mixing, local emission patterns, and the influence of meteorological conditions — it is necessary to incorporate atmospheric variables such as temperature, humidity, wind speed and direction, pressure, and precipitation. Their inclusion is expected to shift the model from correlational mapping toward explanatory and predictive capability.

Conclusions

The conducted research confirmed the effectiveness of the Random Forest algorithm in correlational modeling of PM2.5 concentration, based on data obtained from the OpenWeatherMap Air Pollution API. The results obtained – a coefficient of determination R^2 of 0.994 in validation and 0.970 in testing – indicate a high precision of model fit. At the same time, the feature importance analysis showed a clear dominance of the PM10 variable in the decision structure, confirming the strong correlation between these indicators and indicating the limited role of the other variables in the current dataset. From a safety engineering perspective, this model can be treated as a tool to support the monitoring of environmental data integrity and the detection of anomalies in measurement systems. The high consistency between predicted and observed values allows the model to be used to identify deviations from typical relationships between PM10 and PM2.5, which may indicate measurement errors, sensor failures, or unusual environmental events. This approach is an important element in the development of preventive environmental safety systems focused on early detection of threats. The limitations of the model result from the lack of meteorological variables and the excessive dependence between input data. Future work should include additional weather parameters (temperature, humidity, wind speed and direction, pressure, precipitation) and sequential time series data, which will enable the construction of models with real predictive power. The introduction of deep learning algorithms, such as LSTM recurrent networks or transformational models (TFT), could allow for modeling the dynamics of changes in air quality and predicting smog episodes in advance. In summary, the presented research is a step towards the use of machine learning methods in environmental safety engineering. The Random Forest-based model can be interpreted as a component of a Safe Environmental Monitoring system, which in the future—after being expanded with additional data sources—could become an integral part of intelligent decision support systems for environmental safety and environmental risk management. Overall, the findings indicate that the Random Forest model functions mainly as a correlational estimator of PM2.5 based on PM10, rather than a forecasting tool. Achieving true predictive performance will require including meteorological drivers and temporal dynamics in future versions of the model.