Improving Public Health Policies with Indoor Air Quality Predictive Models

Ariel Isaac Posada Barrera
Universidad Popular Autónoma del Estado de Puebla (UPAEP),
Departamento de ingenierías,
Facultad de Tecnologías de Información y Ciencia de Datos (FTIyCD)
Puebla, Mexico
arielisaac.posada@upaep.edu.mx

Laura Margarita Rodríguez Peralta
Universidad Popular Autónoma del Estado de Puebla (UPAEP),
Departamento de ingenierías,
Facultad de Tecnologías de Información y Ciencia de Datos (FTIyCD)
Puebla, Mexico
lauramargarita.rodriguez01@upaep.mx

Éldman de Oliveira Nunes
Centro Universitário SENAI/CIMATEC
Bahía, Brasil
eldman.nunes@gmail.com

Paulo Nazareno Maia Sampaio
Universidade Salvador (UNIFACS)
Bahía, Brasil
pnms.funchal@gmail.com

Fabian Leonardo Cuesta Astudillo
Universidad Politécnica Salesiana,
Cuenca, Ecuador,
Universidad Popular Autónoma del Estado de Puebla (UPAEP),
Departamento de ingenierías,
Facultad de Tecnologías de Información y Ciencia de Datos (FTIyCD)
Puebla, Mexico
fabian.cuesta@upaep.edu.mx

***Abstract*— Indoor air quality is important for public health. This study was designed to develop predictive models focusing on indoor air quality, specifically targeting levels of CO2, TVOC, PM2.5, and PM10. We implemented and trained Machine Learning Models—Regression Forest Model and Gradient-Boosted Tree Model—using a dataset from the states of Puebla and Morelos in Mexico. The dataset incorporated various environmental variables, including pollutant levels, temperature, relative humidity, population density, and ventilation characteristics, all of which were found to significantly influence the presence of indoor air contaminants. These findings are instrumental in formulating policies to mitigate poor indoor air quality. Moreover, the study suggests that it is feasible to predict when contaminants will reach harmful levels by monitoring changes in these variables.**

Keywords*— *IoT, Big Data, Machine learning, Monitoring, Sick buildings syndrome.

I. Introduction

In today's fast-paced world, a significant portion of daily activities occurs indoors. With individuals spending a considerable amount of time inside, the importance of understanding and ensuring the quality of the air we breathe cannot be overstated. This necessity highlights the imperative to monitor and manage contaminant levels diligently to safeguard the health of individuals in these frequently occupied spaces.

Building upon the data analyzed in previous studies [1], the current research narrows its focus on specific factors. These include ventilation patterns, the concentration of individuals in a space, and prevailing environmental conditions.

Delving into pollutants such as CO2, TVOC, PM2.5, and PM10, our proposal aims to elucidate their interactions and correlations with indoor environmental data. One of the significant insights gleaned from the findings is the considerable influence that human activities exert on indoor air quality.

Indoor Air Quality (IAQ) plays a crucial role in ensuring the health and well-being of individuals residing or working within buildings. With the relentless march of urbanization and the trend of spending more time indoors, the need for accurate prediction and vigilant monitoring of indoor air pollutants has never been more urgent.

Past studies in this domain have typically relied on data from the same pollutants to predict future conditions. These research efforts have been directed at forecasting the Indoor Air Quality Index by utilizing historical indoor pollutant data, meteorological data (such as outdoor temperature and relative humidity), and information on sources of contaminants. The culmination of these studies has provided us with the Air Quality Index and valuable forecasts of future indoor environmental conditions.

In Table I, recent studies that utilized historical data are outlined. These studies have employed historical data pertaining

to pollutant concentrations as part of their inputs, along with the Air Quality Index, to conduct their analyses and predictions.

TABLE I. RECENT STUDIES UTILIZING HISTORICAL DATA AS INPUTS, AND PREDICTING FUTURE AIR QUALITY INDEX

Study	*Model inputs*	*Model outputs*
[2]	Historical data including CO2	Air quality future index and future range of CO2
[3]	Meteorological data including temperature, humidity, wind speed, and wind direction, air quality data from AQMS, structural information and air exchange rate of classrooms	Real time air quality estimation, and CO2 and PM2.5 concentration
[4]	CO, CO2, NO2, O3, Formaldehyde, TVOC, PM and air humidity and temperature	Pollution levels and real time air quality index
[5]	Bioaerosol, PM, UV-induced fluorescence	Bioaerosol and PM estimates, future concentration predictions
[6]	PM2.5, Latent variables (PLS)	Air quality index, and PM2.5 range

In the comprehensive systematic review conducted in [7], several key studies were identified. These studies, distinctively without reliance on historical data, focused on real-time prediction of indoor air pollutants. They utilized indoor environmental data for their predictive models but did not incorporate information related to ventilation patterns or door openings in their analyses.

Table II provides a detailed overview of the inputs and outputs of the models used in these studies.

TABLE II. POLLUTANT LEVELS PREDICTION STUDIES ACCORDING TO [7]

Study	*Model inputs*	*Model outputs*
[8]	Humidity and temperature	Pollutants concentration
[9]	Humidity and temperature	Pollutants concentration
[10]	Environmental data	PM2.5 levels

Building upon the pollutant prediction studies outlined in the previous table, the current research, conducted in Mexico, embarks on implementing a machine learning solution to predict the presence of contaminants that elevate risks in indoor environments. The study utilizes variables selected from the systematic literature review conducted in [1], aiming to generate new knowledge pertinent to the Mexican context.

Our research is driven by the objective of establishing clear, actionable guidelines that enhance ventilation practices and ensure environmental safety in indoor environments. This endeavor is pivotal for fostering healthier indoor settings and mitigating risks associated with respiratory conditions and Sick Building Syndrome (SBS).

To substantiate its findings, our study employs Regression Forest and Gradient-Boosted Tree models, incorporating the selected variables from study [1]. This methodology provides robust evidence to underpin the research's conclusions. The data-intensive components of the research are adeptly managed through the Big Data platform, PySpark, with further details expanded upon in the subsequent sections.

Our proposal had the following research aims: first, developing models for predicting contaminants in indoor spaces in Mexico; second, verifying the existence of codependency between pollutants and various factors, including ambient temperature, relative humidity, the number of people per cubic meter, and the type of ventilation. This verification process is crucial as it underpins the justification for policies related to temperature control, relative humidity, ventilation, and social distancing within indoor environments.

To provide a comprehensive understanding of the complexities involved in predicting indoor air quality, the paper is organized as follows: Section II, "Collected Data," elaborates on the data collection methods and types of variables recorded; Section III, "Machine Learning," focuses on the algorithms and evaluation metrics; Section IV, "Lessons Learned," offers key insights into the role of human activities and environmental conditions; Section V, "Conclusions," summarizes the study's findings and implications; and Section VI, "Future Works," discusses potential avenues for further research.

II. COLLECTED DATA

The data was collected with the Databot sensor described in [11] and the PMS5003 sensor described in [12]. The Databot sensor provided information on environmental variables such as relative humidity, temperature, presence of CO2, and Total Volatile Organic Compounds (TVOC). In contrast, the PMS5003 sensor was responsible for capturing data on particulate matter, specifically PM2.5 and PM10 concentrations. Some data sets were obtained using the air quality monitor [13], which already contains such data (temperature, relative humidity, CO2, TVOC, PM2.5, PM10). Additional metadata was also recorded, including the type of ventilation system in use, whether the door was open or closed during the data collection period, the number of people present, and the volume of the room. These measurements were conducted in various locations across the Mexican states of Puebla and Morelos.

Before taking the readings, a comparison was made between the measurement values obtained from the Databot and PMS5003 systems and those from the Air Quality Monitor IAQM-128W. This comparison was conducted simultaneously, considering different environmental conditions. It was observed that there were no significant differences in the variable values between the systems.

A dataset was obtained with a total of 74,826 unique data scenarios, with no missing column values, in measurements taken over periods ranging from 1 to 6 hours at the locations using both technologies.

The dataset's graphics indicate a mean temperature of approximately 24.59°C. The average levels of CO2 and TVOC

were recorded at 513.37 ppm and 239.21, respectively, while the mean humidity level was 16.61%. The dataset shows PM2.5 and PM10 levels having mean values of 9.42 and 11.35, respectively. The people density, calculated based on the dimensions of the indoor spaces and logs of people entering and exiting the rooms, had an average value of 0.11. For the one-hot encoded variables, approximately 42% of the data points had Normal ventilation activated, around 29% had Natural ventilation, and in about 69% of the cases, the Door was open. The data does not exhibit normal distributions but encompasses a variety of values, serving as a robust dataset for training models. Further details on the dataset's distribution and range for each variable can be observed in Figure 1 below.

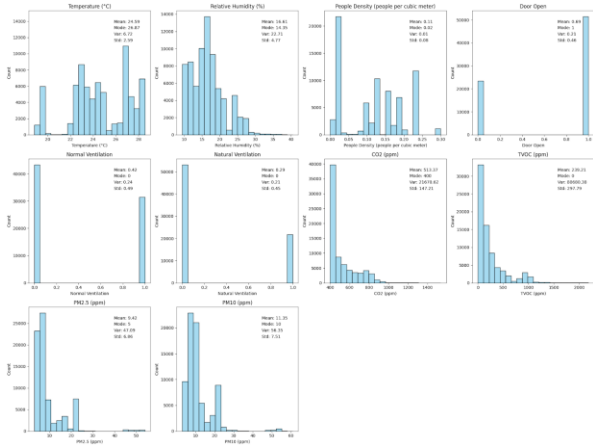

Fig. 1. Histograms of the obtained dataset, with measures of central tendency and variability.

The collected data was utilized in machine learning models to forecast the presence of indoor contaminants. This approach leverages the comprehensive dataset to provide actionable insights for improving indoor air quality.

III. Machine Learning

Building on the data outlined in the previous section, this study employs two main types of models: Random Forests and Gradient-Boosted Tree Models. Random Forests are comprised of an ensemble of decision trees, each trained on different subsets of the data. This approach aims to enhance the model's accuracy and robustness by leveraging the diversity among the trees, as referenced in [14].

In contrast, Gradient-Boosted Tree Models operate differently. These models sequentially build an ensemble of weak learners, typically decision trees, with each tree in the sequence working to correct the errors made by its predecessor. This sequential improvement process is crucial for enhancing the model's predictive performance, as detailed in [15].

The fundamental differences between Random Forests and Gradient-Boosted Tree Models are rooted in their training approach and objectives. While Random Forests seek to create a diversified set of trees through techniques like bootstrapping and feature selection, Gradient-Boosted Trees are designed to sequentially improve the model by minimizing residual errors. This distinction in methodologies results in unique strengths and applications for each model in predicting indoor air quality based on the dataset described earlier.

The evaluation of the models was conducted using the scores of MAE (Mean Absolute Error), which is the average of the absolute differences between predicted and actual observations, providing a straightforward and less outlier-sensitive metric. MSE (Mean Squared Error) is the average of the squared differences between predicted and actual observations, giving more weight to larger errors and being more sensitive to outliers. RMSE (Root Mean Square Error) is the square root of MSE, also emphasizing larger errors and sensitivity to outliers ([16]).

Additionally, the Coefficient of Determination (R-squared) was used to evaluate the machine learning models. R-squared is a statistical measure that represents the proportion of the variance in the dependent variable that can be explained by the independent variable(s) in a regression model. It was chosen for its ability to provide additional information, such as the goodness-of-fit of the model to the data. A higher R-squared value generally indicates a better fit and, therefore, a more reliable algorithm ([17]).

For all the trained models, the selected input variables consisted of Temperature, Relative Humidity, People Density, Door Open, Normal Ventilation, and Natural Ventilation. In all instances, random sampling was employed, allocating 80% of the data for model training and the remaining 20% for testing.

The machine learning regression models, specifically Random Forest and Gradient-Boosted Tree Model, were implemented in Python with Apache Spark ([18]). These models were trained using the predefined input variables, with the primary focus on predicting CO2, TVOC, PM2.5, and PM10 levels.

To facilitate data interpretation and provide a clear understanding of CO2 levels, statistical normalization was applied with a minimum threshold of 400 and a maximum of 1,534. The detailed results obtained from these models, including the prediction of CO2 levels, are presented in Table III.

TABLE III. Model metrics for CO2 prediction with environmental variables.

Metric	***Regression Forest Model***	***Regression Forest Model***
RMSE	0.089	0.069
MSE	0.008	0.005
MAE	0.067	0.045
R-squared	0.530	0.717

In the results, we can observe a discernible difference in the performance between the Regression Forest Model and the Gradient-Boosted Tree Model. The Gradient-Boosted Tree Model outperformed the Regression Forest Model across all metrics. Specifically, the Gradient-Boosted Tree Model

achieved a lower RMSE of 0.069 compared to the Regression Forest Model's 0.089, indicating more accurate predictions. Similarly, the Gradient-Boosted Tree Model exhibited a lower MSE of 0.005 and MAE of 0.045, compared to the Regression Forest Model's MSE of 0.008 and MAE of 0.067, highlighting its superior precision and reliability in predicting the actual observations. Furthermore, with an R-squared value of 0.717, the Gradient-Boosted Tree Model explained approximately 71.7% of the variance in the dependent variable, showcasing a better fit to the data compared to the Regression Forest Model's R-squared value of 0.530 or 53.0%. These results underscore the effectiveness and reliability of the Gradient-Boosted Tree Model in predicting indoor air quality.

For the case of TVOC, statistical normalization was applied with a minimum value of 0 and a maximum of 2,184. The training was conducted using the same models, yielding the results in Table IV.

TABLE IV. MODEL METRICS FOR TVOC PREDICTION WITH ENVIRONMENTAL VARIABLES.

Metric	***Regression Forest Model***	***Regression Forest Model***
RMSE	0.056	0.045
MSE	0.003	0.002
MAE	0.041	0.028
R-squared	0.828	0.891

The Gradient-Boosted Tree Model again demonstrated superior performance over the Regression Forest Model across all evaluated metrics. The RMSE for the Gradient-Boosted Tree Model was lower at 0.045, compared to the Regression Forest Model's 0.056, suggesting that the former model provided predictions with higher accuracy for TVOC levels. The Gradient-Boosted Tree Model also yielded a smaller MSE of 0.002 and a lower MAE of 0.028, as opposed to the Regression Forest Model's MSE of 0.003 and MAE of 0.041. These figures indicate that the Gradient-Boosted Tree Model was more precise and reliable in its predictions. Moreover, the R-squared value for the Gradient-Boosted Tree Model was 0.891, meaning it explained approximately 89.1% of the variance in the dependent variable, which is significantly higher than the 82.8% explained by the Regression Forest Model. This higher R-squared value signifies a better fit and understanding of the data by the Gradient-Boosted Tree Model, making it a more reliable tool for predicting TVOC levels in indoor environments.

For the case of PM2.5, the machine learning model was trained, with statistical normalization considering the minimum value was 3 and the maximum was 54. The results are presented in Table V.

TABLE V. MODEL METRICS FOR PM2.5 PREDICTION WITH ENVIRONMENTAL VARIABLES.

Metric	***Regression Forest Model***	***Regression Forest Model***
RMSE	0.049	0.028
MSE	0.002	0.001
MAE	0.027	0.011
R-squared	0.865	0.957

In the analysis for PM2.5 levels, the Gradient-Boosted Tree Model continues to exhibit superior performance over the Regression Forest Model in all metrics used for evaluation. The RMSE of the Gradient-Boosted Tree Model is 0.028, which is significantly lower than the 0.049 RMSE of the Regression Forest Model, indicating that the predictions for PM2.5 levels from the Gradient-Boosted Tree Model are more accurate. Additionally, the Gradient-Boosted Tree Model has a lower MSE (0.001) and MAE (0.011) compared to the Regression Forest Model's MSE (0.002) and MAE (0.027), showcasing its greater precision and reliability in predicting PM2.5 levels.

Furthermore, the R-squared value for the Gradient-Boosted Tree Model is exceptionally high at 0.957, explaining 95.7% of the variance in the dependent variable. This is substantially higher than the R-squared value of 0.865 (or 86.5%) for the Regression Forest Model. The higher R-squared value of the Gradient-Boosted Tree Model indicates a better fit to the data, making it a more dependable choice for predicting PM2.5 levels in indoor environments.

For the case of PM10, statistical normalization was applied, considering the minimum value was set at 3 and the maximum at 61, with the subsequent results obtained accordingly.

TABLE VI. MODEL METRICS FOR PM10 PREDICTION WITH ENVIRONMENTAL VARIABLES.

Metric	***Regression Forest Model***	***Regression Forest Model***
RMSE	0.053	0.026
MSE	0.003	0.001
MAE	0.028	0.012
R-squared	0.837	0.959

The results indicate that PM10 yielded the best predictive outcomes among all the pollutants analyzed. Once again, the Gradient-Boosted Tree Model outshone the Regression Forest Model across all evaluation metrics. With an RMSE of 0.026, the Gradient-Boosted Tree Model was more accurate in its predictions compared to the Regression Forest Model, which had an RMSE of 0.053. The Gradient-Boosted Tree Model also recorded a lower MSE (0.001) and MAE (0.012), demonstrating its superior precision and reliability in predicting PM10 levels compared to the Regression Forest Model's MSE (0.003) and MAE (0.028).

Moreover, the R-squared value for the Gradient-Boosted Tree Model was outstandingly high at 0.959, explaining approximately 95.9% of the variance in the dependent variable.

This is a significant improvement over the Regression Forest Model's R-squared value of 0.837 (or 83.7%). The high R-squared value of the Gradient-Boosted Tree Model not only indicates a better fit to the data but also underscores its reliability and effectiveness as a tool for predicting PM10 levels with high accuracy in indoor environments.

IV. Lessons Learned

The study provided insights into predicting indoor air quality, emphasizing the significance of factors such as ventilation, social distancing, relative humidity, and indoor temperature in Mexico.

Metrics indicated that the Regression Forest Model and Gradient-Boosted Tree Model were effective in generating predictions for environmental pollutants CO2, TVOC, PM2.5, and PM10.

One key finding from the study is the opportunity to improve air quality prediction models by taking into account variations in environmental conditions, including ventilation patterns and room occupancy. The correlation between these factors and pollutant concentration was proven in the prediction models. Integrating considerations for proper ventilation and social distancing within indoor spaces is essential for maintaining air quality. Future modeling approaches should incorporate these factors for enhanced accuracy in predictions.

The study also pointed out areas requiring improvement, especially in the prediction of CO2 levels. The lower R-squared score for CO2 prediction indicates a need for refinement in modeling approaches for this pollutant. Nonetheless, the models were effective in predicting levels of TVOC, PM2.5, and PM10. Predicting CO2 levels may require modeling techniques that consider the interaction of various environmental factors, including relative humidity and indoor temperature.

Finally, this research contributed to the existing body of knowledge and offered a model for air quality forecasting in Mexico. The findings are relevant for indoor environments in Mexico, where climate and social practices may uniquely influence indoor air quality. The study serves as a foundation for future research exploring the challenges of air quality prediction in similar settings, underscoring the importance of accounting for local environmental and social factors in developing prediction models.

V. Conclussions

This study utilized machine learning, Big Data, and IoT to predict indoor air pollutants. The result was a system designed for real-time monitoring of indoor air quality, capable of detecting variations in pollutant levels and initiating timely interventions.

The analysis identified a relationship between variables such as temperature, relative humidity, people density, and ventilation characteristics with indoor air pollutant concentrations. These findings underscore the importance of continuous monitoring of these variables to ensure the health and well-being of individuals in indoor spaces and prevent Sick Building Syndrome (SBS).

Moreover, the research goes beyond mere pollutant prediction, illuminating the significant role of factors such as ventilation influencing indoor air quality. It pinpointed specific areas, like the prediction of CO2 levels, that warrant further exploration and refinement.

Additionally, the study provides valuable insights into the correlation between indoor air quality and environmental and occupancy variables, which are crucial for devising effective mitigation and prevention strategies. The importance of maintaining adequate ventilation and adhering to social distancing norms to enhance indoor air quality was also underscored.

Overall, our study furnishes actionable insights that can substantially inform and influence health policies aimed at fostering healthier indoor environments. This is particularly pertinent in the context of Mexico, where these findings can be integral to the formulation of effective policies for improving indoor air quality and promoting healthier and safer living and working environments.

VI. Future Works

Drawing upon the findings of this study, future work will aim to formulate a real-time monitoring system for the prevention of health risks in indoor environments. An additional objective for subsequent research is to refine the predictive algorithm for CO2 levels. Given the unique challenges associated with forecasting CO2, this area presents an opportunity for further advancements in algorithmic techniques.

Conflict of interest

The authors declare that they have no conflicts of interest.

References

[1] Domínguez Portillo, J., Rodríguez Peralta, L. M., Sampaio, P. N. M., & Nunes, E. d. O. (2023). Determining environmental indicators related to the propagation of contagious diseases and health issues: A Systematic Literature Review. In 2023 18th Iberian Conference on Information Systems and Technologies (CISTI) (pp. 1-5). doi:10.23919/CISTI58278.2023.10212010

[2] Putra, J. C. P., Safrilah, & Ihsan, M. (2018). The prediction of indoor air quality in office room using artificial neural network. In AIP Conference Proceedings. doi:10.1063/1.5042896

[3] Sharma, P. K., Mondal, A., Jaiswal, S., Saha, M., Nandi, S., De, T., & Saha, S. (2021). IndoAirSense: A framework for indoor air quality estimation and forecasting. Atmospheric Pollution Research, 12(1), 10–22. doi:10.1016/j.apr.2020.07.027

[4] Baqer, N.S., Albahri, A.S., Mohammed, H.A. et al. (2022). Indoor air quality pollutants predicting approach using unified labelling process-based multi-criteria decision making and machine learning techniques. Telecommun Syst, 81, 591–613. doi:10.1007/s11235-022-00959-2

[5] Lee, J. Y. Y., Miao, Y., Chau, R. L. T., Hernandez, M., & Lee, P. K. H. (2023). Artificial intelligence-based prediction of indoor bioaerosol

concentrations from indoor air quality sensor data. Environment International, 174(107900), 107900. doi:10.1016/j.envint.2023.107900

[6] Wu, Q., Geng, Y., Wang, X., Wang, D., Yoo, C., & Liu, H. (2023). A novel deep learning framework with variational auto-encoder for indoor air quality prediction. Frontiers of Environmental Science & Engineering, 18(1). doi:10.1007/s11783-024-1768-7

[7] Wei, W., Ramalho, O., Malingre, L., Sivanantham, S., Little, J. C., & Mandin, C. (2019). Machine learning and statistical models for predicting indoor air quality. Indoor Air, 29(5), 704–726. doi:10.1111/ina.12580

[8] Skön, J., Johansson, M., Raatikainen, M., Leiviskä, K., & Kolehmainen, M. (2012). Modelling indoor air carbon dioxide (CO2) concentration using neural network. World Academy of Science, Engineering and Technology, International Science Index, 6, 737–741.

[9] Khazaei, B., Shiehbeigi, A., & Haji Molla Ali Kani, A. R. (2019). Modeling indoor air carbon dioxide concentration using artificial neural network. International Journal of Environmental Science and Technology, 16, 729–736.

[10] Das, P., Shrubsole, C., Jones, B., et al. (2014). Using probabilistic sampling-based sensitivity analyses for indoor air quality modelling. Building and Environment, 78, 171–182.

[11] Databot. (2023). Home - databot™ - Real Data, Real Science, Real Fun! Retrieved from https://databot.us.com

[12] The World Air Quality Project. (2023). The Plantower PMS5003 and PMS7003 Air Quality Sensor experiment. Retrieved from https://aqicn.org/sensor/pms5003-7003/

[13] Inkbird. (2023). Wi-Fi 8-in-1 Air Quality Monitor IAQM-128W. Retrieved from https://inkbird.com/products/wi-fi-air-monitor-iaqm-128w

[14] Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324

[15] Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. The Annals of Statistics, 29(5), 1189–1232. doi:10.1214/aos/1013203451

[16] Willmott, C. J., & Matsuura, K. (2005). Advantages of the mean-absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Climate Research, 30, 79-82.

[17] German, G. (2021). Coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE, and RMSE in evaluating regression analysis. PeerJ Computer Science, 7(e623), e623. doi:10.7717/peerj-cs.623

[18] Zaharia, M., Xin, R. S., Wendell, P., Das, T., Armbrust, M., Dave, A., ... & Ghodsi, A. (2016). Apache Spark: A unified engine for big data processing. Communications of the ACM, 59(11), 56-65.