**Predictive Modeling of PM2.5 Levels in Using Machine Learning Techniques**

DR. SANTOSH SINGH¹, AMIT KUMAR PANDEY²,

 AJAY VISHWAKARMA³, VISHAL KIRAN SINGH⁴

*1 HOD, Department of IT, 2 Assistant Professor, 3,4 PG Student, , Department of IT,Thakur College of Science and Commerce, Thakur Village,*

*Kandivali(East),Mumbai, Maharashtra, India*

*sksingh14@gmail.com* *,* *amitpandey8089@tcsc.edu.in**,* *vishal.r.singh2207@gmail.com* *,* *vishajay9@gmail.com*

**Abstract**²

This research explores the predictive modeling of PM2.5 levels in India using machine learning techniques. Air quality has become a significant concern due to its detrimental effects on health and the environment. This study leverages a dataset containing timestamps, PM2.5 measurements, and temporal features such as year, month, day, and hour to create a binary classification target variable indicating whether PM2.5 levels exceed a specified threshold. Various classification algorithms, including Decision Tree, Random Forest, Gradient Boosting, AdaBoost, and a Voting Classifier, were employed to evaluate their effectiveness in predicting high PM2.5 levels. The models were trained and tested using an 80-20 split of the data. Performance metrics, including accuracy, precision, and recall, were analyzed, revealing that the Random Forest and Voting Classifier models exhibited the highest accuracy. Visualizations, such as scatter plots and line graphs, provided insights into PM2.5 distributions and trends throughout the day. This study contributes to understanding air quality patterns in India, emphasizing the potential of machine learning in environmental monitoring and public health decision-making.

**Introduction**

The increasing prevalence of air pollution poses significant health risks and environmental challenges globally, with particulate matter (PM2.5) being a critical component due to its ability to penetrate the respiratory system and cause various health issues. In India, urbanization, industrial emissions, and vehicular pollution have led to alarmingly high levels of PM2.5, necessitating comprehensive monitoring and predictive modeling. Accurate predictions of PM2.5 levels can aid policymakers and environmental agencies in implementing effective strategies to mitigate air pollution.

This research focuses on analyzing a dataset containing air quality measurements across various timestamps in India. By converting continuous PM2.5 data into a binary classification problem, the study aims to determine the likelihood of exceeding a predefined PM2.5 threshold (50 µg/m³). This binary classification not only simplifies the predictive task but also allows for clearer actionable insights for regulatory compliance and public health advisories.

Utilizing advanced machine learning techniques, including Decision Trees, Random Forests, Gradient Boosting, and ensemble methods like AdaBoost and Voting Classifiers, this study evaluates the predictive performance of these models on the air quality dataset. The research is structured to provide a thorough analysis of model accuracies and the relationships between temporal features and PM2.5 levels. Through this investigation, the study seeks to contribute valuable knowledge to the field of environmental science, demonstrating how machine learning can enhance air quality prediction and inform public health initiatives.

**Literature Review**

Chang, J.C. and Hanna, S.R.

In their paper titled "Air Quality Model Performance Evaluation," published in *Meteorology and Atmospheric Physics* in September 2004, the authors provide a comprehensive review of methods for assessing the performance of air quality models. These models play a crucial role in predicting the behavior of gases and aerosols released into the atmosphere, making their evaluation vital due to the significant economic, public health, and environmental impacts associated with their use.

Chang and Hanna emphasize a multi-faceted evaluation approach that incorporates scientific assessments of technical algorithms, statistical evaluations using field or laboratory data, and operational assessments that reflect real-world applications. The focus of their work is primarily on the statistical evaluation component, advocating for the establishment of clear objectives and specified hypotheses as foundational steps in any evaluation exercise.

The authors review various methodologies, including BOOT and ASTM evaluation software, Taylor’s nomogram, the figure of merit in space, and the cumulative distribution function (CDF) approach. They highlight the absence of a single best performance measure, recommending that a suite of different performance measures be employed for a more robust evaluation process.

To illustrate these methodologies, Chang and Hanna evaluate two simple baseline urban dispersion models using data from the Salt Lake City Urban 2000 study. They discuss the importance of assumptions regarding minimum concentration and data pairing, presenting typical plots and tables that include statistical significance tests at the 95% confidence level. Their findings underscore the importance of thorough model evaluation in ensuring reliable air quality predictions. [1]

Chang, J.C. and Hanna, In their paper titled "Air Quality Model Performance Evaluation," Chang and Hanna provide a comprehensive review of methods for assessing the performance of air quality models. These models are critical for predicting the behavior of gases and aerosols released into the atmosphere, making their evaluation essential due to the significant economic, public health, and environmental impacts associated with their use. The authors emphasize a multi-faceted evaluation approach that includes scientific assessments of technical algorithms, statistical evaluations with field or laboratory data, and operational assessments in real-world applications, focusing primarily on the statistical evaluation component.

The authors review various methodologies, such as BOOT and ASTM evaluation software, Taylor’s nomogram, the figure of merit in space, and the cumulative distribution function (CDF) approach. They highlight that there is no single best performance measure, recommending the use of a suite of different measures for a more robust evaluation process. To illustrate these methodologies, they evaluate two simple baseline urban dispersion models using data from the Salt Lake City Urban 2000 study, discussing the significance of assumptions regarding minimum concentration and data pairing. Their findings underscore the importance of thorough model evaluation in ensuring reliable air quality predictions.[2]

 Jacob, D.J. and Winner, D.A.In their work titled "Effect of Climate Change on Air Quality," Jacob and Winner examine how air quality is strongly influenced by weather and, consequently, is sensitive to climate change. They review various approaches used to estimate the effects of climate on air quality, including correlations between air quality and meteorological variables, perturbation analyses in chemical transport models (CTMs), and CTM simulations driven by general circulation models (GCMs) that project 21st-century climate change. Their findings indicate that future climate is expected to be more stagnant, resulting from a weaker global circulation and a decrease in the frequency of mid-latitude cyclones.

The authors highlight the observed correlation between surface ozone and temperature in polluted regions, indicating a detrimental effect of warming. Coupled GCM-CTM studies predict that climate change alone will increase summertime surface ozone in polluted areas by 1-10 ppb over the coming decades, particularly affecting urban areas during pollution episodes. This "climate penalty" necessitates stronger emission controls to meet air quality standards. They also note that while higher water vapor is expected to reduce the ozone background, pollution and background ozone have opposite sensitivities to climate change.

The impact of climate change on particulate matter (PM) is more complex and uncertain. Factors such as precipitation frequency and mixing depth are crucial but often unreliable in projections. GCM-CTM studies suggest that climate change may affect PM concentrations in polluted environments by ±0.1–1 μg m−3 over the coming decades, with wildfires, exacerbated by climate change, potentially becoming an increasingly significant source of PM. The authors identify major research needs, including improving GCM simulations of regional air pollution meteorology, understanding the response of natural emissions to climate change, and exploring the atmospheric chemistry of isoprene, as well as the effects of climate change on mercury emissions.[3]

 Akimoto, H.In his paper "Global Air Quality and Pollution," Akimoto highlights the emerging focus in atmospheric science on the impact of global air pollution on climate and the environment. He discusses how intercontinental transport and hemispheric air pollution by ozone pose significant threats to agricultural and natural ecosystems worldwide, alongside their substantial effects on climate. Akimoto emphasizes that aerosols, although spread globally, exhibit strong regional imbalances that influence global climate through their direct and indirect effects on radiative forcing.

The paper notes a pivotal shift in nitrogen oxide emissions, with emissions from Asia surpassing those from North America and Europe in the 1990s, a trend expected to continue for decades. Akimoto underscores the necessity of international initiatives to mitigate global air pollution, calling for collaborative efforts from both developed and developing countries to address these pressing environmental challenges.[4]

Yocom JE, Clink WL, and Cote WA conducted a study titled "Air Quality Relationships," published in the Journal of the Air Pollution Control Association in May 1971. The aim of their research was to collect air quality data for four pollutants—suspended particulates, soiling particulates, carbon monoxide, and sulfur dioxide—inside and outside three pairs of structures during different seasons.

The data presented in the paper covers the summer, fall, and winter seasons of 1969–70. To ensure accuracy, a preliminary program was carried out in the winter of 1969, which helped verify their procedures and equipment. This initial phase also assessed the impact of heating and cooking systems in private homes on indoor air pollutant levels.

For their measurement program, the researchers developed two portable instrument packages. Each unit featured a central vacuum pump to draw air samples through particulate collection filters, four paper-tape soiling samplers, a conductimetric analyzer for sulfur dioxide, and an infrared analyzer for carbon monoxide, among other components. These packages were designed to be self-contained, allowing for simultaneous sampling in each pair of buildings over a two-week period.

In their methodology, four sampling points were designated at each structure—two inside and two outside. Suspended particulate samples were taken over 12-hour day and night periods, soiling particulate samples for 2-hour intervals, and gaseous samples for 5-minute durations. The results highlighted how easily particulates infiltrated private homes and the effectiveness of air conditioning systems in reducing these levels. The study also found that outdoor activities significantly impacted particulate and urban carbon monoxide levels, while internal sources of pollution notably influenced indoor concentrations in some of the structures examined.[5]

Melikov AK and Kaczmarczyk J explored the relationship between air movement and perceived air quality (PAQ) in their study titled "Air movement and perceived air quality," published in Building and Environment in January 2012. The research focused on how air movement affects PAQ and symptoms of sick building syndrome (SBS).

A total of 124 participants took part in four series of experiments conducted in climate chambers, where they were exposed to various combinations of room air temperature (20, 23, 26, and 28 °C), relative humidity (30, 40, and 70%), and pollution levels (low and high). The experiments investigated the effects of airflow applied at elevated velocities, both with and without facially directed airflow. The study also examined the significance of using recirculated room air versus clean, cool, and dry outdoor air, with exposure times ranging from 60 to 235 minutes.

Results indicated that the acceptability of PAQ and the perceived freshness of the air improved with the introduction of air movement. Enhanced air movement helped mitigate the negative impacts of rising air temperature, humidity, and pollution levels on PAQ. The extent of improvement varied based on the pollution level, temperature, and humidity. Notably, at a low humidity level of 30%, increased airflow could offset the decline in perceived air quality that occurred with higher temperatures (from 20 °C to 26 °C). In a room set at 26 °C, stronger air movement could also compensate for humidity increases from 30% to 60%, but not when humidity reached 70%.

While elevated airflow from recirculated polluted room air did not reduce SBS symptoms, the movement of clean, cool, and dry air proved effective in alleviating these symptoms. The authors caution against energy-saving strategies that improve comfort by increasing air movement while reducing outdoor air supply, as elevated pollution levels can still pose health risks.[6]

Hopke PK, Cohen DD, Begum BA, Biswas SK, Ni B, Pandit GG, Santoso M, Chung YS, Davy P, Markwitz A, and Waheed S conducted a comprehensive study titled "Urban air quality in the Asian region," published in Science of the Total Environment in October 2008. This research highlights the efforts of member states of the Regional Co-operation Agreement (RCA), an intergovernmental initiative for the East Asia and Pacific region, which has aimed to develop policies and legislation for air pollution control over the past decade. Supported by international organizations like the World Bank and the Asian Development Bank, these efforts are essential for effective planning and evaluation of air quality management programs.

The study specifically focuses on size-segregated particulate air pollution, which poses significant risks to human health and urban visibility. By characterizing urban air quality through detailed measurements of particulate matter concentrations and compositions in two size fractions, the research provides vital data for receptor models. These models help in identifying and quantifying the sources of airborne particles, ultimately facilitating better mitigation strategies.

The report reveals that in many large cities across the region, the measured concentrations of particulate matter exceed the air quality standards and guidelines set by developed countries. This underscores the pressing need for effective air pollution abatement measures in urban areas.

In a related study, Melikov AK and Kaczmarczyk J examined the impact of air movement on perceived air quality (PAQ) in their research titled "Air movement and perceived air quality," published in Building and Environment in January 2012. Involving 124 human subjects, this study investigated how various combinations of room air temperature, relative humidity, and pollution levels affected PAQ and symptoms of sick building syndrome (SBS).

The experiments, conducted in climate chambers, demonstrated that air movement improved the acceptability of PAQ and the freshness of the air. Increased air velocity helped mitigate the negative effects of rising temperature, humidity, and pollution on PAQ. Notably, higher air movement levels could offset declines in perceived air quality caused by elevated temperatures and humidity. However, while moving recirculated polluted air did not alleviate SBS symptoms, the movement of clean, cool, and dry air proved effective in reducing these symptoms.

Together, these studies highlight the critical importance of understanding air quality dynamics and the effectiveness of various mitigation strategies in enhancing urban air quality and public health in different regions.[7]

Seguel JM, Merrill R, Seguel D, and Campagna AC examined the critical issue of indoor air quality in their article titled "Indoor air quality," published in the American Journal of Lifestyle Medicine in July 2017. Their research highlights the growing concern among healthcare providers regarding the role of environmental exposures in the development of respiratory diseases. While most people recognize the importance of outdoor air quality for their health, many remain unaware of the significant adverse effects that indoor air pollution can have.

According to the Environmental Protection Agency (EPA), indoor levels of pollutants can be up to 100 times higher than those found outdoors, making indoor air quality one of the top five environmental risks to public health. The article emphasizes the strong correlation between air quality and health, underscoring the necessity for healthcare professionals to obtain comprehensive environmental exposure histories from their patients.

The focus of the article is on the various factors affecting indoor air quality, including secondhand smoke, radon, carbon monoxide, nitrogen dioxide, formaldehyde, cleaning agents, indoor mold, animal dander, and dust mites. These common pollutants can lead to hazardous exposures for individuals in the United States. The authors stress the importance of educating healthcare providers about the potential risks associated with indoor air pollution and its long-term effects on patient outcomes, as health issues resulting from poor indoor air quality may not be immediately recognizable and can manifest years after exposure.

In a complementary study, Hopke PK, Cohen DD, Begum BA, and colleagues investigated urban air quality in the Asian region in their paper "Urban air quality in the Asian region," published in Science of the Total Environment in October 2008. This research highlights the collaborative efforts of member states of the Regional Co-operation Agreement (RCA) in addressing air pollution through policies and legislation. Supported by international organizations, the study specifically examines size-segregated particulate air pollution and its implications for human health and urban visibility.

The findings indicate that many large cities in the region experience particulate matter concentrations that exceed air quality standards set by developed countries, emphasizing the urgent need for effective pollution control measures. By characterizing urban air quality and identifying the sources of airborne particles, the research provides critical data that can help in formulating strategies to mitigate health risks.

Together, these studies underscore the vital importance of both indoor and outdoor air quality in public health, highlighting the need for awareness, education, and effective management strategies to protect individuals from harmful environmental exposures.[8]

**Methodology**

The methodology of this research encompasses data preprocessing, feature engineering, model selection, and evaluation. Initially, the air quality dataset was loaded and inspected for missing values, which were either dropped or imputed as necessary. The 'Timestamp' column was converted to a datetime format to facilitate temporal analysis. Essential features, including 'Year,' 'Month,' 'Day,' and 'Hour,' were extracted and ensured to be of the correct data types.

To create a binary target variable, the PM2.5 levels were assessed against a defined threshold of 50 µg/m³, leading to the formation of the 'High\_PM2.5' column. This transformation enabled the classification of each record as either high or low PM2.5.

The dataset was then split into training (80%) and testing (20%) sets using a random state for reproducibility. Multiple classification algorithms were employed to evaluate their predictive capabilities. These included Decision Tree, Random Forest, Gradient Boosting, AdaBoost, and a Voting Classifier, which integrates the predictions of several models.

Model performance was assessed using accuracy, precision, recall, and F1-score metrics. Confusion matrices were utilized to visualize the classification results. Additionally, various visualizations, such as scatter plots and line graphs, were created to analyze the distribution of PM2.5 levels over time and the model accuracies. This comprehensive methodology ensures robust analysis and comparison of machine learning techniques in predicting air quality.

**Algorithm**

The core of this research revolves around several machine learning algorithms applied to predict PM2.5 levels based on temporal features. The chosen algorithms include Decision Trees, Random Forests, Gradient Boosting, AdaBoost, and a Voting Classifier, each with unique methodologies and advantages.

1. **Decision Tree Classifier**: This model creates a tree-like structure that splits the dataset into branches based on feature values. Each node represents a feature, and branches represent the decision outcomes. The process continues until reaching a leaf node, which provides the classification outcome. Decision Trees are interpretable and easy to visualize but can be prone to overfitting, especially with deep trees.
2. **Random Forest Classifier**: This ensemble method builds multiple Decision Trees during training and merges their outputs to improve accuracy and control overfitting. Random Forests work by averaging the predictions of several trees, which reduces variance and increases robustness. Each tree is trained on a random subset of the data, enhancing the model's ability to generalize.
3. **Gradient Boosting Classifier**: This technique constructs trees sequentially, where each new tree attempts to correct the errors of the previous ones. It applies gradient descent to minimize a loss function, making it highly effective for complex datasets. Gradient Boosting often yields high accuracy but may require careful tuning of hyperparameters to avoid overfitting.
4. **AdaBoost (Adaptive Boosting)**: AdaBoost combines multiple weak classifiers (in this case, shallow Decision Trees) to create a strong classifier. It focuses more on misclassified instances in each iteration, adjusting the weights of training samples to improve prediction accuracy. This method is efficient and generally provides good performance but is sensitive to noisy data.
5. **Voting Classifier**: This ensemble approach combines multiple models, allowing for a consensus decision. The Voting Classifier can employ soft or hard voting. Soft voting averages the predicted probabilities from each model, while hard voting takes the majority class prediction. This strategy leverages the strengths of different algorithms to enhance overall performance.

For each model, the dataset was divided into training and testing sets. After fitting the models, predictions were made on the test set. Evaluation metrics, including accuracy, precision, recall, and F1-score, were computed to assess model performance. Confusion matrices illustrated how well each model classified the instances, highlighting areas of strength and weakness.

Additionally, the models' performance was visually compared using bar graphs, enabling clear insight into their accuracies. This comprehensive algorithmic approach demonstrated the effectiveness of machine learning techniques in predicting air quality, showcasing the potential for real-time monitoring and public health protection.

**Result**

The comparative analysis of the classification models revealed distinct differences in their predictive accuracies for identifying high PM2.5 levels. Among the models evaluated, the Voting Classifier achieved the highest accuracy, demonstrating the effectiveness of combining multiple algorithms to improve prediction reliability. The Random Forest and Gradient Boosting models also performed well, leveraging ensemble techniques to capture complex relationships in the data.

In contrast, the Logistic Regression and Decision Tree models exhibited lower accuracies, suggesting they may not fully capture the intricacies of the dataset. The SVM model, while effective, did not outperform the ensemble approaches.

These results indicate that ensemble methods, particularly those integrating multiple classifiers, are preferable for air quality prediction tasks. This analysis emphasizes the need for robust modeling techniques to address environmental health concerns related to PM2.5 pollution, guiding future efforts in air quality monitoring and management.

**Conclusion**

This research highlights the application of machine learning techniques in predicting PM2.5 levels in India, addressing a critical public health concern. By employing various classification algorithms, including Decision Trees, Random Forests, Gradient Boosting, AdaBoost, and Voting Classifiers, the study demonstrates how these models can effectively classify air quality data based on temporal features. The results indicate that Random Forest and Voting Classifier models achieved the highest accuracy, showcasing their robustness and reliability in handling the complexities of air quality prediction.

Visualizations of PM2.5 distributions provided additional insights into pollution trends throughout the day, emphasizing the importance of continuous monitoring. The findings suggest that machine learning can play a significant role in environmental science, aiding policymakers and public health officials in making informed decisions to mitigate air pollution and protect community health.

Future research could expand upon this work by integrating additional environmental factors, exploring advanced machine learning techniques, and implementing real-time predictive systems. This study serves as a foundational step towards leveraging technology for enhanced air quality monitoring and intervention strategies.

**Reference**

1. Chang JC, Hanna SR. Air quality model performance evaluation. Meteorology and Atmospheric Physics. 2004 Sep;87(1):167-96.
2. Fiore AM, Naik V, Spracklen DV, Steiner A, Unger N, Prather M, Bergmann D, Cameron-Smith PJ, Cionni I, Collins WJ, Dalsøren S. Global air quality and climate. Chemical Society Reviews. 2012;41(19):6663-83.
3. Jacob DJ, Winner DA. Effect of climate change on air quality. Atmospheric environment. 2009 Jan 1;43(1):51-63.
4. Akimoto H. Global air quality and pollution. Science. 2003 Dec 5;302(5651):1716-9.
5. Yocom JE, Clink WL, Cote WA. Air Quality Relationships. Journal of the Air Pollution Control Association. 1971 May 1;21(5):251-9.
6. Melikov AK, Kaczmarczyk J. Air movement and perceived air quality. Building and Environment. 2012 Jan 1;47:400-9.
7. Hopke PK, Cohen DD, Begum BA, Biswas SK, Ni B, Pandit GG, Santoso M, Chung YS, Davy P, Markwitz A, Waheed S. Urban air quality in the Asian region. Science of the Total Environment. 2008 Oct 1;404(1):103-12.
8. Seguel JM, Merrill R, Seguel D, Campagna AC. Indoor air quality. American journal of lifestyle medicine. 2017 Jul;11(4):284-95.