IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

PREDICTIVE MODELING OF PM2.5 LEVELS USING MACHINE LEARNING TECHNIQUES

DR. SANTOSH SINGH[1], AMIT KUMAR PANDEY[2],

ANIKET NANHELAL SAROJ[3], SURYA PRATAP YADAV[4]

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

Abstract

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.

Index Terms- PM 2.5, Machine Learning, Air Quality Prediction, Classification Algorithms, Environmental Monitoring

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 modelling. Accurate predictions of PM2.5 levels can aid policymakers and environmental agencies in implementing effective strategies to mitigate air pollution.

This research focuses on analysing 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.

NEED OF THE STUDY

Air pollution, particularly PM2.5, has become a significant environmental and public health concern worldwide, especially in rapidly urbanizing countries like India. High concentrations of PM2.5 contribute to respiratory diseases, cardiovascular issues, and reduced air quality, making effective monitoring and prediction crucial for mitigating its adverse effects.

Traditional air quality monitoring methods rely on physical sensor networks, which, while accurate, are often costly, limited in coverage, and lack real-time predictive capabilities. To overcome these challenges, there is a growing need for data-driven approaches that can enhance air quality monitoring and enable proactive decision-making.

This project is essential because:

1. Public Health Protection: Accurate PM2.5 predictions can help in issuing early warnings to reduce exposure to hazardous air conditions.
2. Policy and Urban Planning: Predictive insights allow government agencies and environmental organizations to implement effective pollution control measures.
3. Technological Advancement: Leveraging machine learning provides a scalable and efficient alternative to traditional air quality assessment methods.
4. Data-Driven Decision Making: By converting continuous PM2.5 data into a binary classification problem, the study ensures clear and actionable insights for regulatory compliance.
5. Enhanced Predictive Capabilities: Using advanced ML models like Decision Trees, Random Forests, Gradient Boosting, and ensemble methods improves the reliability of air quality forecasts.

This research contributes to environmental science and machine learning applications, providing a robust framework for real-time air quality prediction and supporting public health initiatives.

Literature Review

Air pollution has been a significant concern globally, with numerous studies focusing on the prediction, monitoring, and evaluation of air quality. Various research efforts have explored different methodologies, including machine learning, climate modelling, and statistical assessments, to enhance the accuracy of air pollution predictions.

Air Quality Model Performance Evaluation

Chang and Hanna (2004) conducted a comprehensive review of air quality model evaluations, emphasizing statistical, operational, and scientific assessment approaches. Their work highlighted the importance of multi-faceted evaluation methods, including BOOT and ASTM evaluation software, Taylor's nomogram, and the cumulative distribution function (CDF) approach. The study underscores the necessity of employing multiple performance measures for robust model evaluation. Their analysis of urban dispersion models using Salt Lake City Urban 2000 study data demonstrated the significance of defining clear objectives and hypotheses in model evaluation.

Effect of Climate Change on Air Quality

Jacob and Winner (2009) investigated the impact of climate change on air quality. Their study found a strong correlation between surface ozone levels and temperature, predicting increased summertime surface ozone due to climate change. They identified factors such as decreased mid-latitude cyclone frequency and increased

stagnation, which may lead to deteriorating air quality. The study also suggested that climate change could influence particulate matter (PM) levels through alterations in precipitation and mixing depth, though projections remain uncertain. Key research areas include improving climate models for regional air pollution and understanding how natural emissions respond to climatic changes.

Global Air Quality and Pollution

Akimoto (2003) examined the global impact of air pollution, particularly intercontinental transport and hemispheric ozone pollution. His research emphasized the role of nitrogen oxide emissions, noting that Asia had surpassed North America and Europe in emissions since the 1990s. The study highlighted the necessity for international collaborations to mitigate global air pollution's effects on climate and ecosystems.

Air Quality Relationships

Yocom et al. (1971) investigated the relationship between indoor and outdoor air pollutants, analyzing suspended particulates, carbon monoxide, and sulfur dioxide levels. Their study revealed the significant impact of outdoor pollution on indoor air quality, with external activities influencing urban carbon monoxide levels. The effectiveness of air conditioning systems in reducing indoor pollution was also highlighted, emphasizing the need for improved air filtration and ventilation strategies.

Air Movement and Perceived Air Quality

Melikov and Kaczmarczyk (2012) explored the effects of air movement on perceived air quality (PAQ) and symptoms of sick building syndrome (SBS). Their study, involving climate chamber experiments, demonstrated that increased airflow improves PAQ by mitigating the adverse effects of high temperature, humidity, and pollution. They noted that while recirculated polluted air does not alleviate SBS symptoms, clean, cool, and dry air significantly improves air quality perceptions. The findings caution against reducing outdoor air supply as a cost-saving measure, as it can increase pollution exposure risks.

Urban Air Quality in the Asian Region

Hopke et al. (2008) examined air quality management policies across Asia, focusing on particulate matter pollution. The study revealed that many large Asian cities experience PM concentrations exceeding developed countries' air quality standards. It emphasized the role of receptor models in identifying pollution sources and the urgent need for effective mitigation strategies. The research supports international initiatives for improving air quality monitoring and policy formulation.

Indoor Air Quality and Public Health

Seguel et al. (2017) highlighted the critical importance of indoor air quality, noting that indoor pollutants can reach concentrations up to 100 times higher than outdoor levels. The study discussed various indoor pollution sources, including secondhand smoke, radon, carbon monoxide, and volatile organic compounds. It emphasized the need for awareness and improved ventilation systems to mitigate long-term health risks associated with poor indoor air quality.

Conclusion

The reviewed literature underscores the significance of predictive modelling in air quality assessment. Various studies have demonstrated the effectiveness of statistical and machine learning approaches in evaluating and forecasting PM levels. The findings highlight the role of climate change, urbanization, and policy interventions in shaping air quality trends. As air pollution continues to be a pressing global challenge, integrating advanced modelling techniques and interdisciplinary collaborations is essential for developing effective air quality management strategies.

RESEARCH 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.

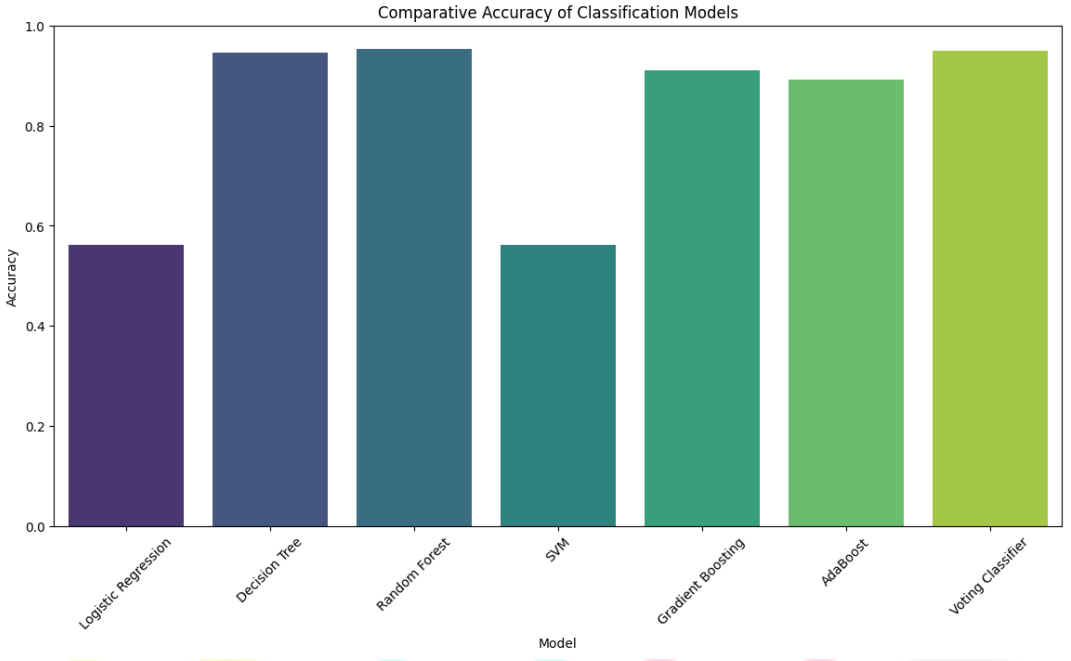
Additionally, the models' performance was visually compared using bar graphs, enabling clear insight into their accuracies. This comprehensive algorithmic approach demonstrated the effectiveness of machine learning techniques in predicting air quality, showcasing the potential for real-time monitoring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

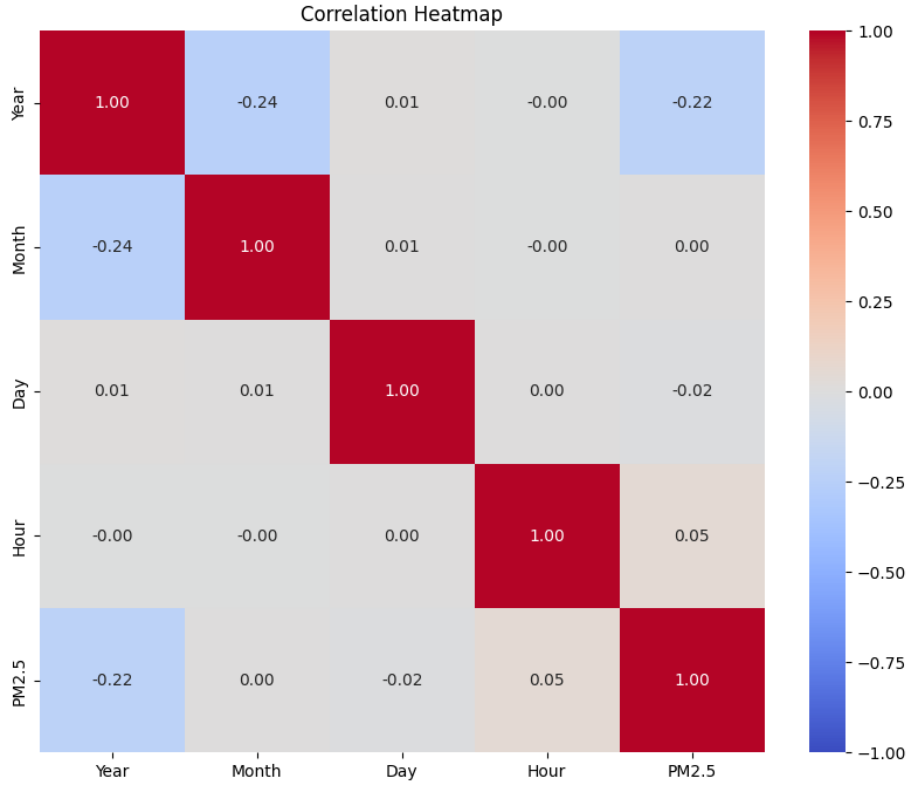
RESULTS AND DISCUSSION

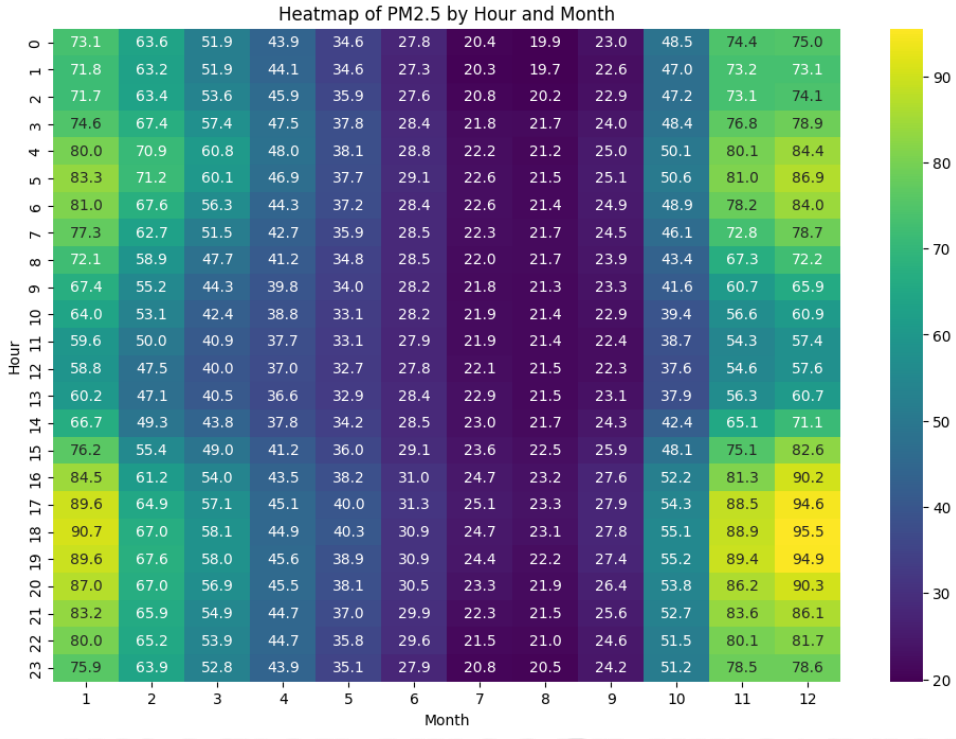
The comparative analysis of the classification models revealed distinct differences in their predictive accuracies for identifying high PM2.5 levels. Among the models evaluated, the Voting Classifier achieved the highest accuracy, demonstrating the effectiveness of combining multiple algorithms to improve 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.







Acknowledgement

We sincerely thank **Amit Kumar Pandey** for his guidance and **Dr. Santosh Singh**, HOD, for his support. We are grateful to **Thakur College of Science and Commerce** for providing resources and encouragement. Our teamwork made this project possible.

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.

REFERENCES

[1] Ali, A. 2001.Macroeconomic variables as common pervasive risk factors and the empirical content of the Arbitrage Pricing Theory. Journal of Empirical finance, 5(3): 221–240.

[2] Basu, S. 1997. The Investment Performance of Common Stocks in Relation to their Price to Earnings Ratio: A Test of the Efficient Markets Hypothesis. Journal of Finance, 33(3): 663-682.

[3] Bhatti, U. and Hanif. M. 2010. Validity of Capital Assets Pricing Model.Evidence from KSE-Pakistan.European Journal of Economics, Finance and Administrative Science, 3 (20).