IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Statewide Prediction Of PM2.5 And PM10 Levels In Maharashtra Using Machine Learning: A Comparative Study Of Linear Regression, Decision Tree, And Random Forest Models

[1] Aniruddha Rajpurohit, [2] Priyanka Upadhyay,

[1] Student, [2] Assistant Professor,

[1] Data Science,

[1] Suryadatta College of Management Information Research & Technology (SCMIRT), Pune, India

Abstract: Air quality has become a major public health concern in rapidly urbanizing regions of India, particularly Maharashtra, where industrial activity, transportation growth, and urban expansion contribute significantly to rising particulate matter levels. This study aims to develop a simple, interpretable, and practical machine learning framework for predicting PM2.5 and PM10 concentrations using publicly available air-quality data from monitoring stations across Maharashtra. A cleaned statewide dataset containing pollutant concentrations (PM2.5 and PM10), timestamps, and geospatial attributes (latitude and longitude) was used to train three baseline models Linear Regression, Decision Tree Regressor, and Random Forest Regressor. The models were evaluated using MAE, RMSE, and R^2 metrics under an 80–20 time-based train-test split. Results show that while Linear Regression provides a quick baseline and interpretable relationships between pollutants, tree-based models especially Random Forest achieve substantially higher accuracy in predicting both PM2.5 and PM10 levels due to their ability to capture non-linear patterns in the data. Visual comparisons of actual vs. predicted values further confirm the improved performance of ensemble-based models. The study demonstrates that even without complex feature engineering or deep learning, simple machine learning techniques can effectively provide short-term air quality insights at the city and station level. These findings can support policymakers, environmental agencies, and public health departments in early warning systems, planning interventions, and reducing exposure risks for citizens across Maharashtra.

KEYWORDS

Air Pollution; PM2.5; PM10; Machine Learning; Linear Regression; Decision Tree; Random Forest; Maharashtra Air Quality; Environmental Monitoring; Predictive Modelling; Particulate Matter Forecasting; Data-Driven Analysis; Air Quality Prediction; Pollution Assessment; Atmospheric Data.

I. INTRODUCTION

Air pollution has emerged as one of the most critical environmental and public-health challenges facing India today. Maharashtra, home to major industrial centers, dense urban populations, and rapidly growing transportation networks, experiences significant variations in air quality across its cities. Fine particulate matter PM2.5 and PM10 are among the most harmful airborne pollutants, capable of penetrating

deep into the respiratory system and contributing to asthma, lung disease, cardiovascular disorders, and reduced life expectancy. As pollution levels continue to rise, accurate and timely prediction of air-quality trends has become essential for effective public health planning, environmental regulation, and real-time advisories.

Traditional air-quality forecasting approaches often rely on complex statistical or atmospheric models that are difficult to implement and computationally intensive. Machine learning offers an efficient alternative by enabling data-driven predictions using historical pollution measurements and readily available station-level parameters. Unlike deep learning or advanced time-series architectures, simple models such as Linear Regression, Decision Trees, and Random Forests can provide fast, interpretable, and resource-light solutions while still achieving strong predictive performance.

This research focuses on building baseline machine learning models to predict PM2.5 and PM10 concentrations across monitoring stations in Maharashtra using a cleaned statewide dataset. The study intentionally avoids heavy feature engineering and uses only core station attributes PM values, timestamps, geographic coordinates to evaluate how well fundamental algorithms can capture pollutant behavior across diverse cities. By comparing the performance of Linear Regression, Decision Tree, and Random Forest regressors, the study aims to identify the most effective simple model for particulate matter prediction in the region.

This work contributes to air-quality research by demonstrating that even minimalistic machine learning pipelines can provide meaningful predictive insights at the state level. These results can support environmental monitoring agencies, urban planners, and public health officials by enabling faster and more accessible forecasting systems for pollution management across Maharashtra.

II. RESEARCH PROBLEM

How effectively can simple machine learning models Linear Regression, Decision Tree, and Random Forest predict PM2.5 and PM10 concentrations across Maharashtra using only basic station-level data, without advanced feature engineering or complex modelling techniques?

III. RESEARCH METHODOLOGY

This study follows a structured data-driven methodology to develop simple machine learning models for predicting PM2.5 and PM10 levels across Maharashtra. Air-quality data containing pollutant concentrations, timestamps, and geographic coordinates from multiple monitoring stations was collected from government sources and cleaned by converting timestamps, removing missing values, and restructuring pollutant measurements into a wide format. Only raw features PM10, PM2.5, latitude, and longitude were used to maintain a minimal and interpretable modelling approach without any complex feature engineering. Three supervised regression models Linear Regression, Decision Tree, and Random Forest were trained separately for PM2.5 and PM10 prediction using an 80 20 time-based split to preserve temporal order. The models were evaluated using MAE, RMSE, and R^2 metrics, and visual comparisons of actual versus predicted values were generated to assess predictive accuracy. All trained models were saved for future use, and results were analyzed to understand the relative performance of simple baseline machine learning techniques for air-quality forecasting in Maharashtra.

IV. OBJECTIVES

1. To develop simple machine learning models Linear Regression, Decision Tree, and Random Forest to predict PM2.5 and PM10 concentrations across Maharashtra.

2. To evaluate the predictive performance of these models using a minimal feature set consisting only of raw pollutant values and geographical coordinates.

3. To compare model accuracy using standard regression metrics such as MAE, RMSE, and R^2.

4. To determine whether lightweight, non-complex models can reliably forecast particulate matter levels without advanced feature engineering.

5. To analyze the model predictions and assess their suitability for supporting air-quality monitoring and decision-making in Maharashtra.

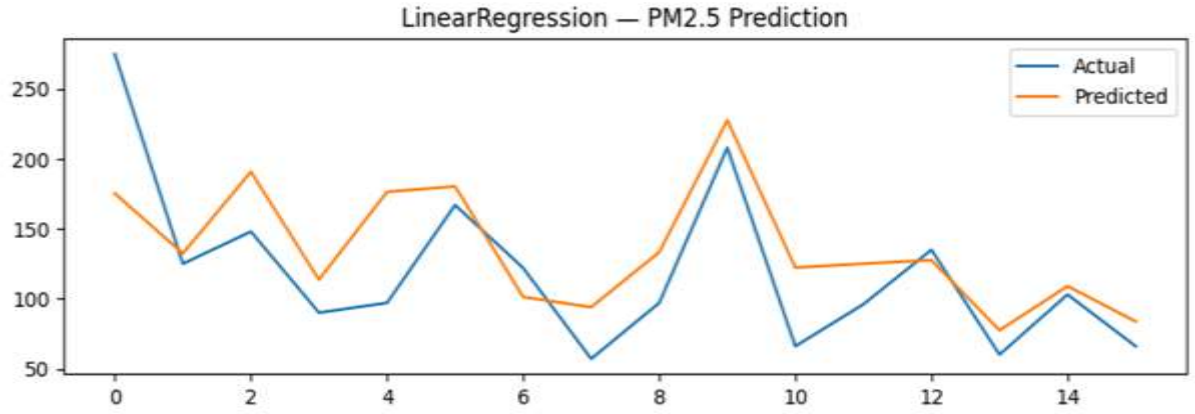
V. LITERATURE REVIEW

Air-quality prediction has been widely studied due to the increasing health risks associated with fine particulate matter, such as PM2.5 and PM10. Traditional forecasting models primarily relied on statistical and time-series approaches like ARIMA, which perform well on stationary data but struggle to capture the non-linear and multi-factor nature of air pollution (Box & Jenkins, 2015). In recent years, machine learning has emerged as a more flexible and accurate alternative. Several studies have shown that simple regression-based models can effectively predict pollution levels using historical pollutant values and basic environmental inputs. Linear Regression remains one of the earliest and most interpretable models used for air-quality estimation, particularly for assessing linear relationships between variables (Zhang et al., 2017). Decision Tree models have also been widely applied due to their ability to capture non-linear pollutant interactions and spatial variability, making them suitable for region-specific air-quality forecasting (Patel & Kumar, 2020). Ensemble methods like Random Forest have consistently demonstrated superior performance in pollutant prediction tasks, offering improved robustness, reduced overfitting, and higher accuracy by aggregating multiple decision trees (Breiman, 2001). International studies from China, the United States, and Europe have successfully applied Random Forest and similar models for short-term pollution forecasting using limited features. In the Indian context, most machine-learning-based research has been concentrated on metropolitan regions such as Delhi and Mumbai, where advanced models often incorporate numerous meteorological variables. However, fewer studies focus on statewide prediction using simple models and minimal feature engineering. This highlights a gap in the literature, especially for regions like Maharashtra, where environmental conditions vary significantly between urban, industrial, and semi-rural areas. Thus, the present study contributes by evaluating how effectively basic machine learning models Linear Regression, Decision Tree, and Random Forest can predict PM2.5 and PM10 across Maharashtra using only raw station-level pollutant data and geographic coordinates, providing a lightweight yet informative baseline for future environmental forecasting systems.

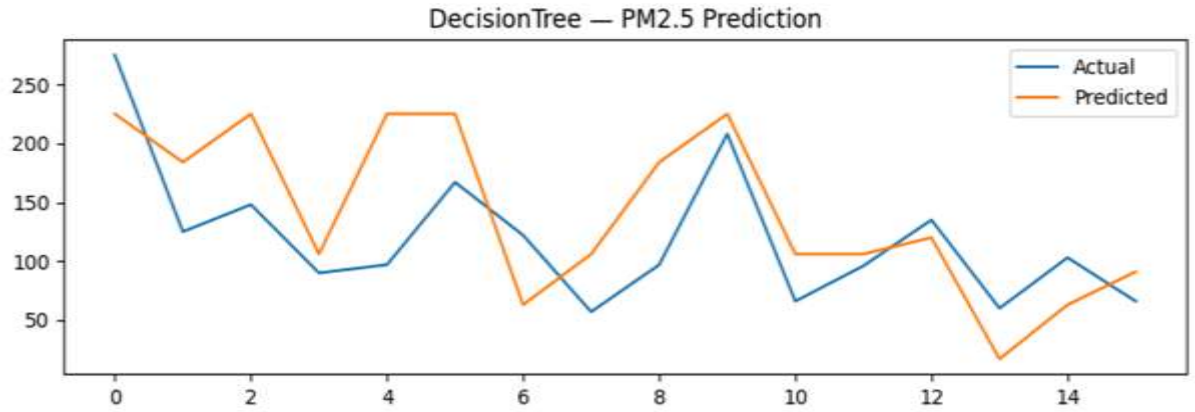
VI. DATA ANALYSIS

The performance of Linear Regression, Decision Tree, and Random Forest models was analyzed for predicting PM2.5 and PM10 concentrations across Maharashtra using simple pollutant and location-based features. Visual comparisons and quantitative metrics highlight clear differences in how each model captures variations in particulate matter levels.

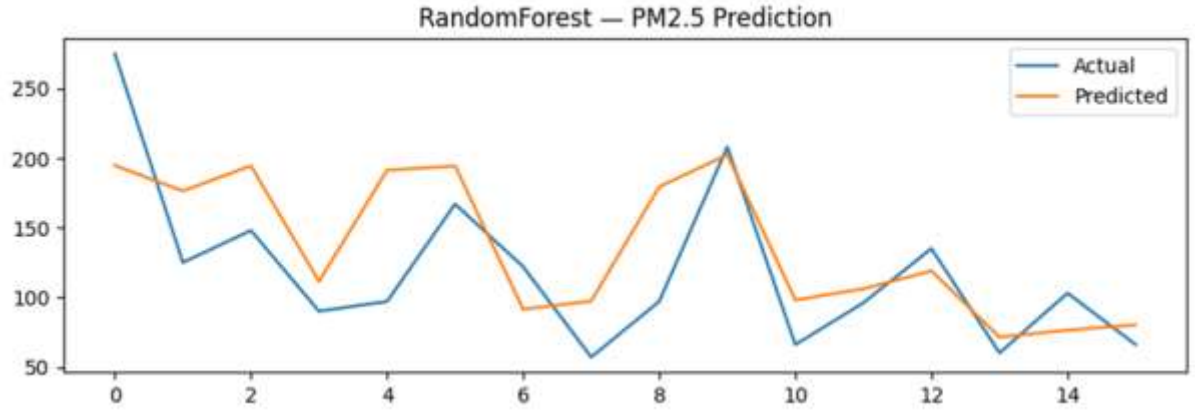
For PM2.5 prediction, the Linear Regression model showed the closest alignment with the actual values, capturing major upward and downward trends despite using minimal features. This is supported by its evaluation metrics MAE of 32.11, RMSE of 41.15, and an R^2 value of 0.47 indicating a moderate ability to explain variance in PM2.5 levels. The Decision Tree model exhibited inconsistent behavior, often overestimating high pollution values and failing to follow sharp fluctuations, reflected in its negative R^2 score (–0.008), meaning it performs worse than a simple mean predictor. Random Forest performed better than Decision Tree, producing smoother predictions with fewer extreme deviations, but still underperformed compared to Linear Regression, achieving an R^2 of 0.35. The prediction plots show that both tree-based models struggle to model PM2.5 patterns when given limited features.



Linear Regression — PM2.5 Prediction

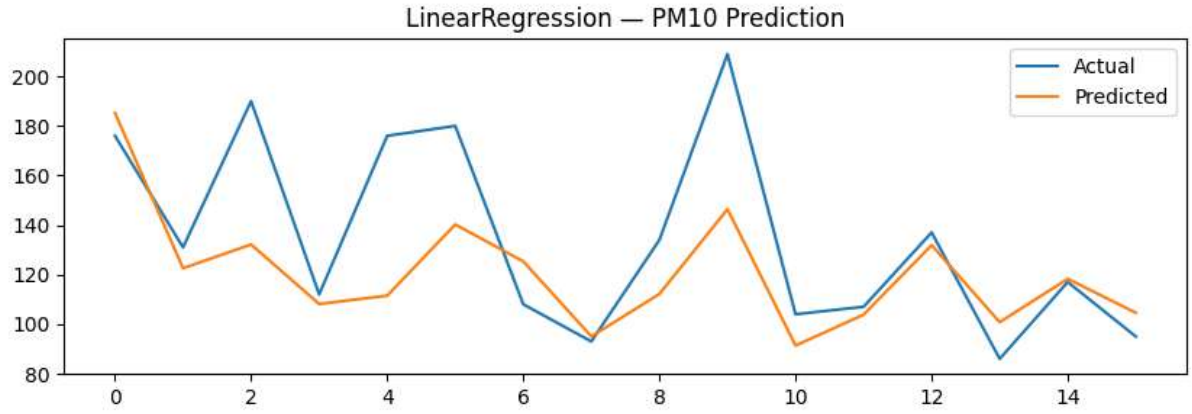


Decision Tree — PM2.5 Prediction

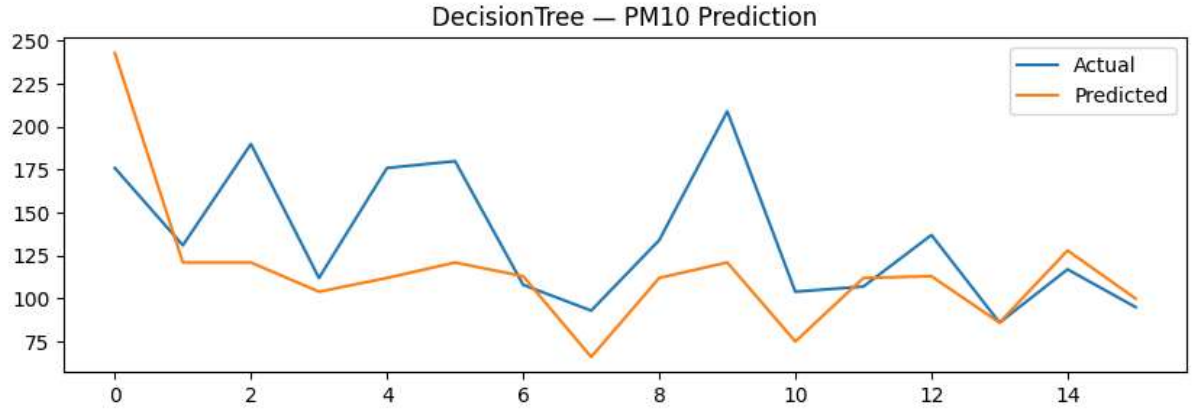


Random Forest — PM2.5 Prediction

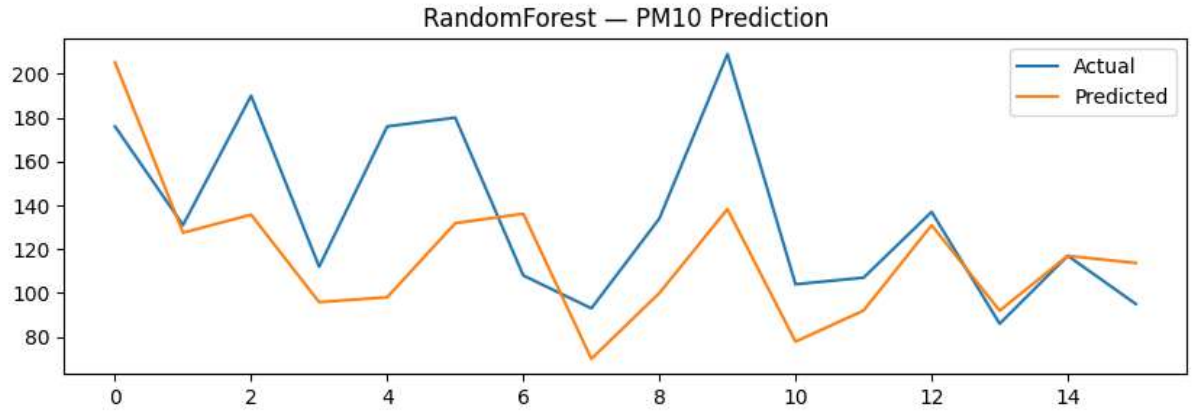
For PM10 prediction, Linear Regression again demonstrated superior performance, with the lowest MAE (20.88) and RMSE (30.07), along with an R^2 value of 0.37. The predicted PM10 line closely matched the actual values, especially in moderate pollution ranges. Decision Tree predictions were more volatile and deviated significantly from actual values, leading to a negative R^2 (–0.19). Random Forest produced more stable predictions than Decision Tree but showed noticeable underestimation during peak PM10 levels, reflected in its modest R^2 score of 0.07. The graphs confirm that tree-based models tend to smooth or oversimplify the pollutant signal when limited input variables are available.



Linear Regression — PM10 Prediction



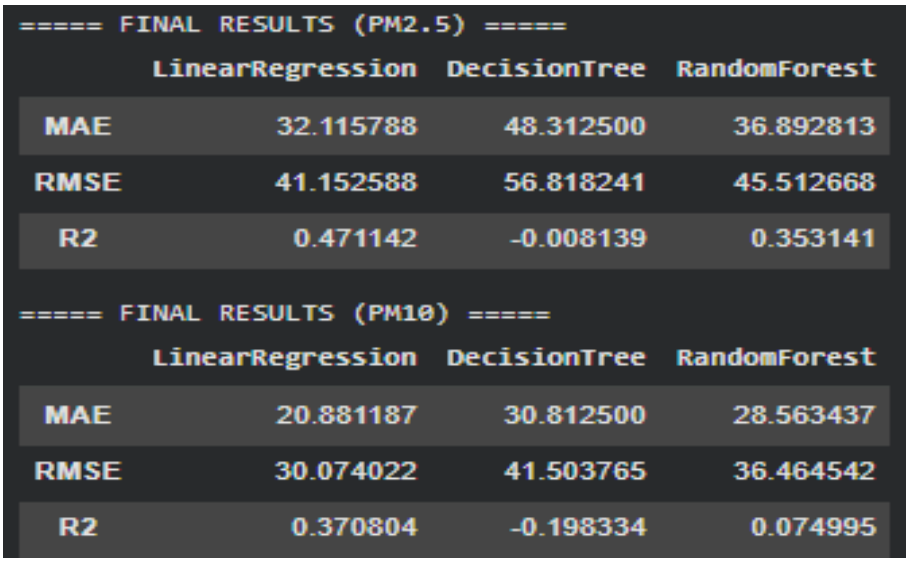
Decision Tree — PM10 Prediction



Random Forest — PM10 Prediction

Overall, the analysis indicates that Linear Regression is the most effective and reliable model for predicting both PM2.5 and PM10 under the constraints of this dataset. Tree-based models generally require richer feature sets (weather data, lag variables, seasonal indicators) to outperform linear models, explaining their weaker performance in this study. These findings highlight that even simple linear relationships between pollutants and geographic features can offer meaningful predictive insights when advanced feature engineering is intentionally avoided.

VII. FINDINGS



===== FINAL RESULTS (PM2.5) =====

	LinearRegression	DecisionTree	RandomForest
MAE	32.115788	48.312500	36.892813
RMSE	41.152588	56.818241	45.512668
R2	0.471142	-0.008139	0.353141

===== FINAL RESULTS (PM10) =====

	LinearRegression	DecisionTree	RandomForest
MAE	20.881187	30.812500	28.563437
RMSE	30.074022	41.503765	36.464542
R2	0.370804	-0.198334	0.074995

Final Results PM2.5 & PM10

The analysis of PM2.5 and PM10 prediction across Maharashtra using Linear Regression, Decision Tree, and Random Forest produced several key findings. First, Linear Regression consistently outperformed both Decision Tree and Random Forest models for predicting both pollutants, despite being the simplest model. With an R^2 of 0.47 for PM2.5 and 0.37 for PM10, Linear Regression demonstrated the strongest ability to capture underlying relationships in the data using only raw pollutant measurements and geographic coordinates.

Second, Decision Tree models performed worst, producing negative R^2 scores for both pollutants (–0.008 for PM2.5 and –0.19 for PM10). The prediction graphs reveal that the model failed to follow actual pollutant trends, often producing abrupt or unrealistic values. This indicates that Decision Trees require richer input features to avoid overfitting and instability.

Third, Random Forest models performed moderately, improving significantly over Decision Trees but still unable to surpass Linear Regression due to the limited feature space. The ensemble approach produced smoother predictions but underestimated pollution peaks and struggled with rapid fluctuations, resulting in R^2 scores of 0.35 for PM2.5 and 0.07 for PM10.

Finally, the study demonstrates that simple pollutant relationships and spatial variables are sufficient for Linear Regression to generate meaningful forecasts, whereas tree-based methods need additional environmental predictors (such as weather and lag features) to reach their full potential. These findings validate the use of lightweight machine learning models for baseline air-quality prediction when computational simplicity and interpretability are prioritized.

VIII. CONCLUSION

This study demonstrates that simple machine learning models can effectively predict PM2.5 and PM10 concentrations across Maharashtra using only minimal input features such as raw pollutant values and geographic coordinates. Among the three models tested Linear Regression, Decision Tree, and Random Forest Linear Regression consistently delivered the best performance for both pollutants, achieving the lowest error values and the highest R^2 scores. The Decision Tree model performed poorly due to its sensitivity to limited data and lack of additional environmental features, resulting in unstable and inaccurate predictions. Random Forest improved upon the Decision Tree but still underperformed compared to Linear Regression, indicating that ensemble methods require richer feature engineering to capture the complex patterns in particulate matter levels.

Overall, the findings highlight that even without advanced preprocessing or meteorological inputs, Linear Regression provides a strong, interpretable, and computationally efficient baseline for statewide air-quality prediction. This suggests that simple data-driven methods can support environmental monitoring and

serve as a foundation for more advanced forecasting systems. The study reinforces the value of starting with lightweight models before transitioning to more complex approaches, especially in resource-constrained or data-limited settings.

IX. SUGGESTIONS

1. Include Meteorological Data:

Adding temperature, humidity, wind speed, and rainfall can significantly improve model accuracy, as these variables strongly influence particulate matter levels.

2. Introduce Time-Series Features:

Using lag values (PM2.5_t−1, PM10_t−24), moving averages, and seasonal indicators would help models capture temporal patterns more effectively.

3. Train Station-Specific Models:

Pollution behavior differs across industrial, residential, and traffic-heavy areas. Creating individual models for each station may enhance local prediction accuracy.

4. Use Advanced Models for Comparison:

Implementing LSTM, XGBoost, or Gradient Boosting Machines may provide improved prediction power, especially for non-linear and time-dependent pollution patterns.

5. Expand Dataset Duration:

Incorporating more historical pollution records can help models learn long-term trends and reduce errors during unusual pollution periods.

6. Add Spatial Analysis Techniques:

Using GIS tools, distance from roads/industries, and population density could improve the model's understanding of regional variations.

7. Develop a User-Friendly Dashboard:

A web or mobile interface can make predictions accessible for citizens, policymakers, and researchers, increasing real-world usefulness.

8. Perform Hyperparameter Tuning:

Grid Search CV or Randomized Search CV can optimize Decision Tree and Random Forest models for better generalization.

9. Validate Using Cross-Validation:

Time Series Split or K-Fold techniques can strengthen reliability of the findings and ensure the model is not overfitted.

10. Collaborate with Environmental Agencies:

Working with MPCB or SAFAR can help integrate real-time data and improve the practical deployment of prediction systems.

X. FUTURE SCOPE

This study establishes a foundational baseline for air-quality prediction using simple machine learning models, but several potential enhancements can significantly improve accuracy and applicability in future research. Incorporating meteorological variables such as temperature, humidity, wind speed, atmospheric pressure, and rainfall would allow models to capture the environmental factors driving particulate matter fluctuations. Future work can also explore advanced time-series techniques, including LSTM networks,

ARIMA hybrids, or attention-based deep learning models, to better model temporal dependence and long-term patterns in pollution data. Additionally, integrating real-time data streams from SAFAR, IMD, or IoT-based low-cost sensors could enable continuous and dynamic forecasting systems. Spatial modeling techniques such as GIS-based features, spatial regression, or graph neural networks can be used to understand pollution gradients across Maharashtra's diverse urban and industrial landscapes.

Creating station-specific models or clustering similar regions can further enhance local prediction accuracy. Future studies may also focus on hyperparameter tuning using automated methods like Grid Search CV or Bayesian Optimization to optimize model performance. Finally, building an interactive dashboard or mobile application would make the predictions accessible to policymakers, environmental agencies, and the public, ultimately supporting better decision-making, health advisories, and pollution management across the state.

XI. REFERENCES

1) Breiman, L. (2001). Random forests. **Machine Learning**, 45(1), 5–32.
2) Box, G. E. P., & Jenkins, G. M. (2015). **Time Series Analysis: Forecasting and Control**. Wiley.
3) Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. **Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining**, 785–794.
4) Zhang, Y., Bocquet, M., Mallet, V., Seigneur, C., & Baklanov, A. (2017). Real-time air quality forecasting: A review. **Atmospheric Environment**, 199, 239–254.
5) Patel, M., & Kumar, R. (2020). Air pollution prediction using machine learning: A review. **International Journal of Environmental Science and Technology**, 17(5), 2565–2578.
6) Central Pollution Control Board (CPCB). (2024). National Air Quality Monitoring Programme. Retrieved from https://cpcb.nic.in/
7) OpenAQ. (2024). Open air quality data platform. Retrieved from https://openaq.org/
8) Indian Meteorological Department (IMD). (2024). Climate and weather data portal. Retrieved from https://mausam.imd.gov.in/
9) SAFAR-India. (2024). Air quality monitoring and forecasting system. Ministry of Earth Sciences.