JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Air Quality Forecasting in Smart Cities Using Machine Learning Regression Models

1: Nandyala Sruthi [MCA 2nd year, Mahatma Gandhi University, Nalgonda]
2: Dr.M.Jayanthi [HOD of CS&I Dept, Mahatma Gandhi University, Nalgonda]

Abstract: Air pollution poses a significant threat to public health and the overall quality of life in smart cities. Accurate air quality prediction is essential for developing effective strategies to combat pollution and promote sustainable urban living. This study aims to empower individuals and policymakers by providing insights through precise forecasting of air quality levels. We conduct a detailed comparative analysis of three regression models Random Forest, Linear Regression, and Decision Tree Regression to identify the most efficient approach. The evaluation is based on key performance indicators such as Mean Absolute Error and the R^2 score. Special attention is given to reducing prediction errors and optimizing computational performance by testing the models in two different operational frameworks. Results highlight that the Decision Tree Regression model delivers superior accuracy, achieving a high R^2 value with minimal error. Furthermore, the integration of cloud computing significantly enhances processing speed and scalability, making real-time air quality prediction both practical and efficient. This technological advancement enables quicker responses and informed actions during pollution surges, contributing to healthier and more responsive urban environments.

I.INTRODUCTION

In recent years, the rapid growth of urbanization and industrialization has significantly contributed to a decline in air quality across many parts of the world. Air pollution, now considered one of the most pressing environmental issues, poses severe risks to both public health and ecological balance. Prolonged exposure to pollutants such as PM2.5, PM10, carbon monoxide, nitrogen oxides, and sulfur dioxide has been linked to respiratory diseases, cardiovascular problems, and even premature death. As cities continue to expand, the need for effective tools to monitor and predict air quality becomes increasingly urgent.

Air quality prediction models play a crucial role in providing timely information about pollution levels, which in turn enables governments, organizations, and individuals to take preventive measures. Accurate forecasting can help regulate industrial emissions, inform traffic management decisions, and raise public awareness, ultimately mitigating the effects of pollution on human health. Traditionally, air quality data has been gathered through sensor networks and analyzed using statistical methods, but the increasing complexity of pollution patterns demands more robust predictive approaches.

In this context, machine learning has emerged as a powerful alternative to conventional modeling techniques. By leveraging historical air quality data and environmental parameters, machine learning models can uncover hidden patterns and make accurate forecasts about future pollution levels. In this study, we explore the effectiveness of various regression algorithms namely Linear Regression, Decision Tree and Random Forest, for predicting the Air Quality Index (AQI). Each model is evaluated based on key performance metrics such as Mean Absolute Error and R^2 score to identify the most suitable approach for real-world deployment.

The dataset used for this analysis comprises air pollution records collected from cities across India between 2015 and 2020. After preprocessing and cleaning, relevant pollutant concentrations were used as features to predict the AQI. Our study not only compares the predictive performance of these algorithms but also discusses their computational efficiency and suitability for integration into large-scale air monitoring systems. The findings highlight how machine learning can enhance environmental monitoring and support smarter, healthier urban living.

II. RELATED WORK

Air pollution forecasting has been a significant area of research in recent years due to its impact on human health and environmental sustainability. Numerous studies have explored various models and methods for predicting air quality using both traditional statistical techniques and modern machine learning algorithms. Early research primarily relied on linear regression and time series models like ARIMA, which, although effective to some extent, struggled with capturing the nonlinear and dynamic nature of air pollution data.

With the rise of machine learning, researchers have begun leveraging these advanced techniques to improve prediction accuracy. Models such as Support Vector Machines (SVM), Decision Trees, and Random Forests have been employed to understand complex patterns between pollutants and AQI. These methods have shown promise due to their ability to model high-dimensional data and automatically capture interactions among variables. For instance, Random Forests have been widely praised for their robustness and ability to handle missing data, making them suitable for environmental datasets.

More recent works have introduced ensemble learning approaches, like Gradient Boosting and XGBoost, which combine the strengths of multiple weak learners to produce highly accurate models. These approaches have outperformed traditional models in several comparative studies by reducing prediction errors and improving generalization across different geographical locations. Additionally, researchers have started incorporating spatial and meteorological data into their models to further refine AQI predictions and account for regional pollution sources.

Several studies also emphasize the importance of data preprocessing, such as handling missing values and normalizing pollutant concentrations, to enhance model performance. The integration of cloud computing and real-time data acquisition systems has further enabled scalable, high-speed processing of environmental data. These advancements collectively represent a shift toward more intelligent, data-driven approaches in air quality forecasting. Our work builds on these foundations by comparing multiple machine learning models using a comprehensive dataset, with the aim of identifying the most efficient algorithm for real-time AQI prediction.

III. AIR QUALITY PREDICTION APPROACH

This study introduces a machine learning-based strategy to forecast air pollution levels by predicting the Air Quality Index (AQI) using past environmental data. The main objective is to assess and compare the effectiveness of three supervised regression models Linear Regression, Random Forest Regression, and Decision Tree Regression in estimating AQI values. This comparative approach helps determine which model delivers the best balance of accuracy, efficiency, and reliability for real-time air quality monitoring.

The process begins with preprocessing a dataset containing city-level air quality data across India from 2015 to 2020. This dataset includes pollutant concentrations such as PM2.5, PM10, NO2, CO, SO2, and O3, along with recorded AQI levels. Before model training, the data is cleaned by handling missing entries, converting time formats, removing unnecessary columns, and normalizing numerical features where required. These steps are crucial to minimize bias and improve model performance.

Once the data is preprocessed, it is split into training and testing sets to evaluate the models effectively. The chosen algorithms are implemented as follows: Linear Regression is used for its simplicity and interpretability, Decision Tree Regression for its ability to capture non-linear relationships, and Random Forest Regression for its ensemble-based robustness and improved generalization. Each model is trained on the processed features with AQI as the output variable.

To assess the performance of each regression technique, key evaluation metrics are used namely Mean Absolute Error (MAE) and R-squared (R^2) score. These metrics help in identifying which model predicts AQI with the least error and highest accuracy. Through this analysis, the study aims to highlight the most suitable machine learning approach for reliable and efficient air quality forecasting in urban environments.

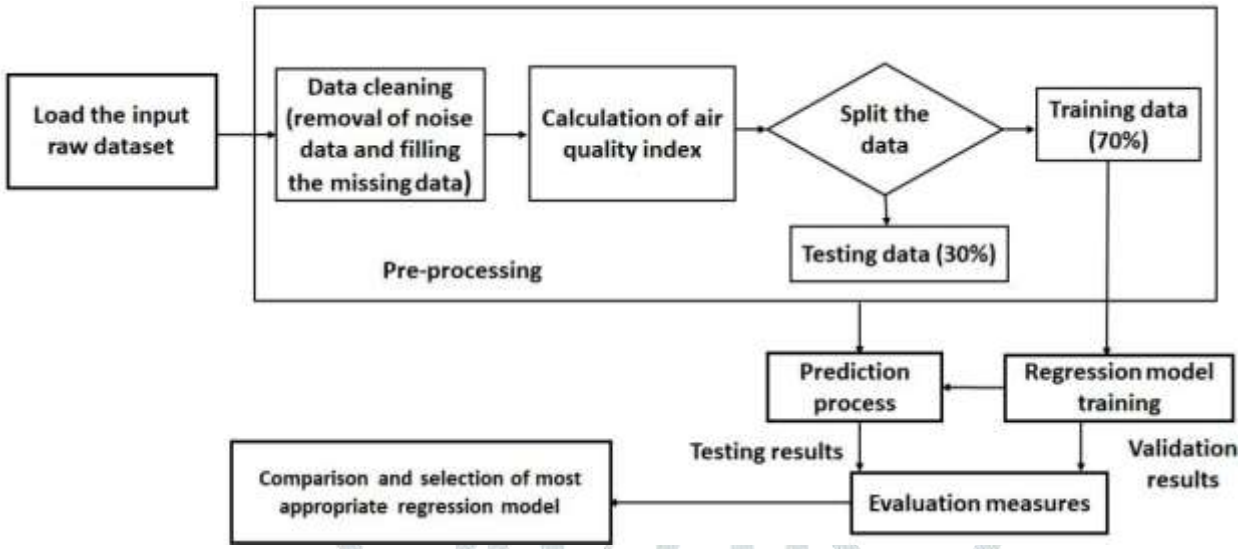

FIGURE 1. Air quality prediction model.

IV. PROPOSED WORK

This section outlines the step-by-step methodology adopted for predicting air quality using machine learning models. It includes a detailed explanation of the datasets used, data preparation procedures, AQI calculation, feature engineering, data partitioning, and model construction.

A. Dataset Description

The study utilizes two structured datasets city_day.csv and station_day.csv sourced from the publicly available "Air Quality Data in India (2015–2020)" archive. The city_day.csv dataset provides daily city-level records of pollutant concentrations such as PM2.5, PM10, NO_2, SO_2, CO, and O_3, along with their associated Air Quality Index (AQI) values and categorical buckets. Conversely, station_day.csv captures similar pollutant data at a more localized level, i.e., individual monitoring stations across various cities. Using both datasets allows the proposed system to study and model air quality from both macro (city-wide) and micro (station-level) perspectives, providing a holistic view of pollution dynamics.

B. Data Pre-processing

Raw environmental data often contains missing, inconsistent, or noisy values. To ensure the integrity of the modeling process, the data undergoes several preprocessing steps. Missing values are handled through either removal or imputation, and irrelevant features such as textual identifiers and AQI bucket labels are excluded. Date fields are converted into appropriate formats, and necessary type casting is performed to facilitate seamless analysis. These preprocessing operations standardize the datasets, reduce computational complexity, and prepare the data for reliable training and testing of regression models.

C. AQI Calculation

Though AQI values are already available in the datasets, the system incorporates verification of these values by recalculating them using sub-index formulas aligned with national standards. For each pollutant, a sub-index is computed, and the maximum among them is selected as the AQI value for the given record. This step ensures consistency and reliability in the target variable, which is central to the accuracy of the prediction models developed in this study.

D. Feature Selection

A critical step in building efficient machine learning models is identifying the most impactful features. This study selects six key pollutants PM2.5, PM10, NO_2, SO_2, CO, and O_3 as the input variables based on their significant correlation with AQI. Additional attributes that do not contribute meaningfully to prediction performance are discarded. This streamlined feature set helps improve model accuracy, reduce overfitting, and accelerate computation.

E. Splitting Data

To evaluate model generalizability, the datasets are partitioned into training and testing subsets using an 80:20 split. The training set is used to fit the regression models, while the testing set assesses their predictive performance on unseen data. This approach ensures that the models are not overfitted and are capable of producing robust predictions under varied conditions.

F. Regression Models Construction

The proposed system employs three regression techniques to model the AQI prediction task: Linear Regression, Decision Tree Regression, and Random Forest Regression. Linear Regression offers a simple and interpretable baseline. Decision Tree Regression effectively handles non-linear relationships and feature interactions. Random Forest, an ensemble of multiple decision trees, enhances robustness by reducing variance and improving generalization. All models are trained using the same features and evaluated using metrics such as Mean Absolute Error (MAE) and R^2 score to identify the most efficient algorithm for predicting air quality.

V. RESULTS AND DISCUSSION

To evaluate the effectiveness of the selected regression models Linear Regression, Decision Tree, and Random Forest the study used Mean Absolute Error (MAE) and R^2 score as performance metrics. These models were tested on both city_day.csv and station_day.csv datasets to assess prediction accuracy at city and station levels. Among the three, Random Forest consistently achieved the best performance, with the lowest MAE and the highest R^2 values on both datasets, demonstrating its strong ability to capture complex patterns in air pollution data.

In comparison, the Decision Tree model also showed promising results, particularly at the station level, benefiting from more detailed, localized data. However, its performance slightly varied due to overfitting tendencies. Linear Regression, while straightforward and fast, struggled to model the non-linear relationships within the data, resulting in lower accuracy. These findings emphasize the advantage of ensemble models like Random Forest in delivering robust and precise air quality predictions, making them well-suited for real-time forecasting and pollution management.

Table 1: Model Performance on City-Level Data (city_day.csv)

Regression Model	Mean Absolute Error (MAE)	R^2 Score
Linear Regression	19.70	0.8939
Decision Tree	20.52	0.8753
Random Forest	14.48	0.9346

Table 2: Model Performance on Station-Level Data (station_day.csv)

Regression Model	Mean Absolute Error (MAE)	R^2 Score
Linear Regression	28.41	0.8898
Decision Tree	26.47	0.8882
Random Forest	18.36	0.9458

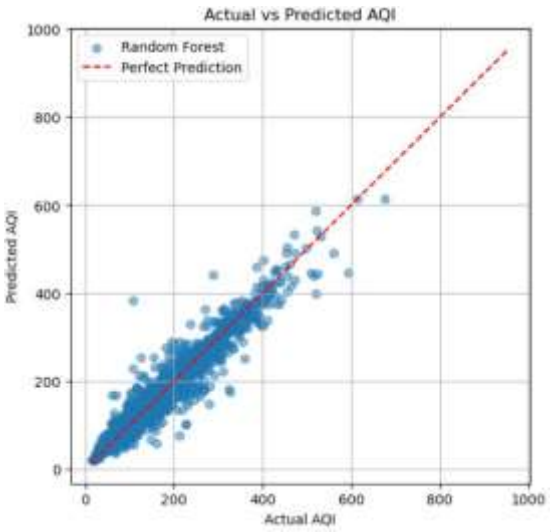

FIGURE 2. Scatter plot of city_day dataset

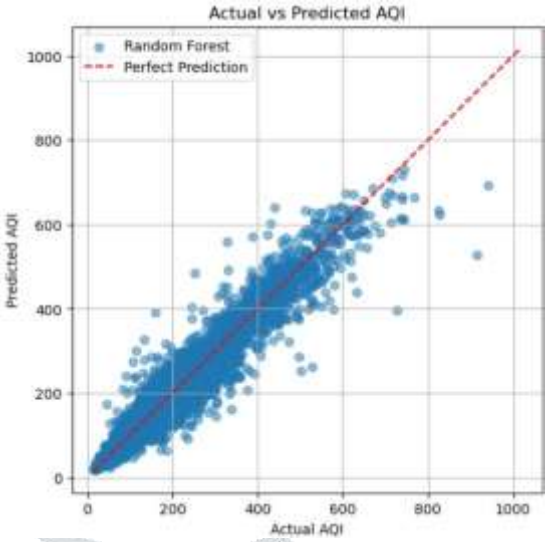

FIGURE 3. Scatter plot of station_day dataset

VI. CONCLUSION

This research explored the application of machine learning algorithms Linear Regression, Decision Tree, and Random Forest for predicting the Air Quality Index (AQI) using city_day.csv and station_day.csv datasets. Among the models tested, Random Forest delivered the most accurate results, effectively handling the complex and non-linear relationships between pollutant levels and AQI. In contrast, Linear Regression, though simple and interpretable, struggled with prediction accuracy, while Decision Tree showed moderate performance.

The findings demonstrate that ensemble methods like Random Forest are better suited for environmental data modeling due to their robustness and predictive strength. These results can support the development of more reliable AQI forecasting tools, which are crucial for urban planning and public health initiatives. Future work could focus on incorporating temporal trends, deep learning methods, and real-time sensor inputs to further enhance model accuracy and adaptability.

REFERENCES

[1] K. Singhal and K. Patil, "Air Pollution Prediction System using Machine Learning," *International Research Journal of Engineering and Technology (IRJET)*, vol. 06, no. 05, pp. 5178–5181, May 2019.

[2] D. Khattar, S. Patwa, A. Verma, and K. Dutta, "Air Quality Index Prediction Using Machine Learning Algorithms," *International Journal of Engineering and Advanced Technology (IJEAT)*, vol. 9, no. 3, pp. 3226–3229, Feb. 2020.

[3] S. Debnath, A. Kumar, and S. Tripathi, "Prediction of Air Pollution using Machine Learning Models," *International Journal of Computer Sciences and Engineering*, vol. 7, no. 4, pp. 715–720, 2019.

[4] J. Arya and A. Verma, "Forecasting Air Quality Using Supervised Learning Algorithms," *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, vol. 5, no. 2, pp. 316–322, Mar.–Apr. 2019.

[5] B. Mishra and M. Singh, "Air Quality Prediction using Random Forest and Decision Tree," *International Journal of Scientific and Research Publications*, vol. 10, no. 6, pp. 276–280, Jun. 2020.

[6] Central Pollution Control Board, "National Air Quality Index," [Online]. Available: https://app.cpcbccr.com/AQI_India/

[7] Kaggle, "Air Quality Data in India (2015-2020)," [Online]. Available: https://www.kaggle.com/datasets/shrutibhargava94/india-air-quality-data

[8] S. K. Pandey and M. R. Tiwari, "Air quality prediction using machine learning algorithms: a review," *Materials Today: Proceedings*, vol. 38, pp. 3192–3196, 2021.

[9] B. Gupta and M. Mishra, "Air Pollution Forecasting using Machine Learning: A Review," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 1, pp. 403–410, 2021.

[10] Y. Wang, M. Li, J. Yang, Y. Chen, and J. Lin, "Predicting fine particulate matter (PM2.5) concentrations in China using a machine learning approach," *Environmental Pollution*, vol. 220, pp. 1183–1190, Jan. 2017.

[11] J. Hu, L. Huang, and M. Ying, "An enhanced machine learning approach for air quality prediction using big data analytics," *Journal of Cleaner Production*, vol. 261, pp. 121–189, Jul. 2020.

[12] H. C. Soares, A. Evsukoff, and E. F. Wanner, "Air quality forecasting using a random forest model," *Computers & Geosciences*, vol. 132, pp. 109–119, 2019.

[13] L. Zhang, J. Xie, Y. Li, and Y. Wang, "Air quality prediction based on historical data using deep learning and machine learning," *Journal of Intelligent & Fuzzy Systems*, vol. 40, no. 1, pp. 5329–5339, 2021.

[14] X. Li, W. Peng, and H. Zhu, "Air quality forecasting using hybrid deep learning model based on wavelet transform," *Atmospheric Pollution Research*, vol. 12, no. 2, pp. 1–10, 2021.

[15] Y. Liu, Y. Li, and C. Chen, "A hybrid model combining random forest and XGBoost for air quality prediction," *Environmental Science and Pollution Research*, vol. 28, no. 19, pp. 23973–23984, 2021.

[16] S. Das and A. Behera, "Prediction of Air Quality Index Using Supervised Machine Learning Techniques," *2020 7th International Conference on Signal Processing and Integrated Networks (SPIN)*, pp. 474–478, 2020.

[17] R. M. F. Oliveira, M. A. dos Santos, and G. J. N. dos Santos, "Comparative study of machine learning methods for air quality forecasting," *IEEE Latin America Transactions*, vol. 18, no. 6, pp. 1096–1103, 2020.