Air Quality Prediction using Machine Learning and Feature Engineering Techniques

Kamali K A
Department of CSE
Rajalakshmi engineering college
Thandalam,, India
220701118@rajalakshmi.edu.in

***Abstract*—Air pollution is a growing phenomenon in urban cities, with serious health consequences and environmental impact. Supporting air quality monitoring and regulation, government agencies employ the Air Quality Index (AQI), a normalized metric of air pollution severity in simple-to-understand terms. This paper proposes a machine learning method to forecast AQI values from a dense environmental data set with pollutant levels (PM2.5, PM10, NO, NO_2, NO_x, NH_3, CO, SO_2, O_3, benzene, toluene, xylene), and metadata such as city and date. The paper explores and compares the performance of four regression models—Linear Regression, Lasso Regression, Ridge Regression, and Decision Tree Regressor—in forecasting AQI values and mapping the values to AQI buckets corresponding to pollution severity. Preprocessing techniques such as missing value imputation, normalization, and outlier removal were employed. Feature engineering used polynomial terms and interaction variables to better capture rich environmental features. Models were trained on 80:20 train-test split and tested with 5-fold cross-validation. Of models tested, Decision Tree Regressor performed best with R^2 of 0.93 and improved ability to predict AQI for a very diverse set of urban environments. This paper points to the promise of interpretable machine learning models for air quality monitoring, offering a scalable, data-driven process for environmental risk mitigation and public health alerts.**

I. Introduction

Air pollution has been a major environmental and public health issue, especially in fast urbanizing areas. Exposure to higher levels of atmospheric pollutants like particulate matter (PM2.5 and PM10), nitrogen oxides (NO, NO_2, NO_x), sulfur dioxide (SO_2), carbon monoxide (CO), ozone (O_3), ammonia (NH_3), and volatile organic compounds (VOCs) like benzene, toluene, and xylene has been linked to respiratory, cardiovascular diseases, and death. Air quality observation and forecasting thus have become central tasks for environmental agencies, city planners, and health authorities.

To make risk management easier and inform the public, government agencies worldwide employ the Air Quality Index (AQI) as a standardized measure that predicts the intensity of air pollution in a particular area and its associated health impacts. The AQI translates complicated pollutant concentration data into a simple-to-understand scale and classifies it into buckets like "Good," "Moderate," "Poor," and "Severe." Precise and timely prediction of the AQI is made possible so that the public can take preventive measures and policymakers can introduce specific interventions to reduce pollution sources.

Traditional air quality prediction techniques normally encompass statistical models and deterministic atmospheric models. They are helpful but typically limited by assumptions of linearity, computational costs, and an inability to define non-linear relations between multiple pollutants and environmental variables. With advancements in machine learning (ML), there are prospects to develop data-driven models to learn non-linear, complex relations from historical data without explicit physical equations. This study employs machine learning techniques in building a prediction model in order to predict AQI values and categorize them into levels of air pollution severity. The provided data set includes samples representing a variety of Indian cities and includes variables such as pollutant concentrations, weather conditions, and date-time variables. Four different algorithms—Linear Regression, Lasso Regression, Ridge Regression, and Decision Tree Regressor—are attempted and compared in relation to high-accuracy predictions for AQI values.

The project pipeline consists of data preprocessing (outlier removal, normalization, and missing value removal), feature engineering (polynomial and interaction feature creation), and cross-validation and hyperparameter tuning during model training. R^2 score, Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE) are the performance metrics used for evaluating the performance of all the models. Model interpretability is also investigated using feature importance analysis, in particular for Decision Tree Regressor, since it provides easy decision-making paths.

Through comparative analysis of various regression model strengths and limitations, this work hopes to establish the most appropriate algorithm for predicting AQI and reveal the most pertinent environmental drivers determining air quality. The resulting outcome is a machine learning algorithm that is understandable and scalable and applicable to feed into real-time air quality warning systems and monitoring and, consequently, support enhanced public health and green urbanization.

II. Literature Review

The increased interest in worsening urban air quality has resulted in the formulation of various predictive models for the prediction of pollutant concentration and Air Quality Index (AQI) values. Improvements in data availability and computational capacity in recent years have made it possible to employ machine learning (ML) methods in this field, which are more accurate and versatile than the older statistical and simulation-based models.

In [1], Support Vector Regression (SVR) and Artificial Neural Networks (ANN) were used to forecast AQI using meteorological variables and pollutant concentrations. While the models were fairly accurate, they were not very interpretable, and performance was highly sensitive to parameter choice. To make the models more interpretable and reduce overfitting, researchers have more and more turned toward regularized regression models. Lasso and Ridge regression were employed in [2] to reduce multicollinearity between predictors while retaining useful features. These models worked well with high-dimensional data but were not quite good enough at detecting non-linear interactions.

Decision tree-family algorithms such as Random Forest and Gradient Boosted Trees have performed well in air quality prediction tasks. A Random Forest was used to train on a dataset of PM2.5, CO, and NO_2 levels and meteorological factors in [3]. The model had an R^2 of greater than 0.90 and identified temperature and wind speed as significant factors. Similarly, [4] employed a Decision Tree Regressor on hourly pollution data and noted its handling of missing values and generation of feature importance scores. The models were also very interpretable, a minimum requirement in application to public health.

Feature engineering has proven to be a key factor in improving prediction precision. In [5], polynomial features and interaction terms were added to capture the non-linear dependence between pollutants and environmental factors. Their experiments best reflected improved regression precision, especially in ensemble models. Domain-specific feature selection was also best explored in [6], where the authors discovered temporal features such as month and hour to have significant impacts on pollutant behavior.

Air quality forecasting data sets are typically plagued with missing values, outliers, and irregular sampling periods. For these, [7] proposed a preprocessing pipeline consisting of imputation, normalization, and outlier detection through Z-score and interquartile range methods. These were discovered to enhance model robustness and generalizability considerably.

In [8], the AQI classification was framed as a multi-class classification problem, utilizing machine learning models to predict AQI bucket labels from raw pollutant measurements. While classification was accurate, regression-based AQI estimation using continuous outputs offered more detail and better performance for downstream analysis.

Overall, previous research confirms that tree-based and regularized regressor machine learning models can actually model complex environmental systems if supplemented with sufficient data preprocessing and feature engineering. However, there is no rigorous comparison of different regression models on a large AQI data set with traditional and engineered features, which the present research study aims to fill.

III. Proposed Methodology

The suggested methodology of air quality prediction is a sequential process involving data preprocessing, feature engineering, model training, and testing on the basis of four machine learning algorithms: Linear Regression, Lasso Regression, Ridge Regression, and Decision Tree Regressor. The workflow steps are as follows:

A. Data Exploration

The dataset comprises air quality measurements of various Indian cities, such as pollutants PM2.5, PM10, NO, NO_2, NO_x, NH_3, CO, SO_2, O_3, benzene, toluene, and xylene, and their corresponding AQI and AQI bucket labels. Missing values, especially in gas measurements such as NO and NH_3, were observed during the initial exploration and were addressed by mean imputation. Outliers were detected by Z-score analysis, in particular, on PM2.5 and PM10 measurements. A correlation heatmap revealed high correlations among PM2.5, PM10, and AQI, which implies their relevance while training the model. The information was also normalized and encoded wherever necessary in order to make it regression algorithm-compatible.

B. Data Preprocessing

The dataset was subjected to systematic preprocessing in order to enhance data quality and make it fit for machine learning models.

1) Missing Value Handling: Few of the features, specifically NO, NO_2, NH_3, and benzene, contained missing values because of sensor malfunctioning or faulty data logging. Missing continuous feature values were filled by the mean of every pollutant over its respective city to maintain spatial significance.

2) Outlier Detection and Removal: Outliers were identified through the Z-score technique. Values with Z-scores greater than ±3 were deemed anomalous and dropped, particularly in PM2.5 and PM10, which often exhibited sensor spikes in industrial areas.

3) Categorical Encoding: The "City" column, a categorical variable, was encoded into numerical format using label encoding to enable model compatibility. Date data was parsed to extract "month" and "day" as extra temporal features.

4) Feature Scaling: To standardize the different ranges of values of pollutants and meteorological variables, Min-Max Scaling was used. It scaled all features to a uniform range of 0 to 1 and facilitated better convergence in regression models such as Lasso and Ridge.

5) Train-Test Split: Post-preprocessing, the dataset was split into 80% training and 20% test using stratified sampling on the AQI bucket to preserve class distribution while evaluating regression.

These steps ensured that the data was clean, consistent, and well-prepared for regression-based AQI prediction.

C. Feature Engineering

To improve the performance of models and to identify complex relationships between variables in the data, some of the feature engineering methods used were:

- **Polynomial Features:** Second-order polynomial features were created for important pollutants like

PM2.5, PM10, and CO to identify non-linear impacts. This enhanced models such as Linear and Ridge Regression to fit curvilinear relationships between the pollutants and AQI more effectively.

- **Interaction Terms:** Multiplicative interaction features were constructed among correlated features, e.g., PM2.5 × Humidity and NO_2 × Temperature. These pairings served to unveil conditional dependencies, enhancing expressiveness of the model.
- **Temporal Features:** The "Date" feature was factorized to reveal features such as month and day, which pick up seasonal and temporal patterns in pollution levels. These features prove particularly valuable for cities experiencing monsoon or winter smog influences.
- **Feature Selection:** Correlation analysis and tree-based model feature importance were employed to detect redundant or weak features. Low-correlating features with AQI, like xylene and toluene in certain locations, were optionally dropped to minimize noise.

D. Model Selection and Training

To predict the Air Quality Index (AQI) from environmental and pollutant data, four supervised regression algorithms were implemented and evaluated:

1) Linear Regression: A baseline model that assumes a linear relationship between input features and the AQI. It provides a simple and interpretable benchmark but may struggle with non-linear data patterns.

2) Lasso Regression: This model adds L1 regularization to Linear Regression, promoting sparsity by shrinking less important feature coefficients to zero. It is useful for feature selection and reduces the risk of overfitting in high-dimensional data.

3) Ridge Regression: Ridge adds L2 regularization, which penalizes large coefficients without eliminating them entirely. It performs well in the presence of multicollinearity and is more stable than Lasso in noisy datasets.

4) Decision Tree Regressor: A non-parametric model that partitions the data space into regions based on feature thresholds. It captures non-linear relationships effectively and provides inherent feature importance scores, making it interpretable and robust for environmental data.

Training Strategy:

- All models were implemented using the scikit-learn library in Python.
- The dataset was split into 80% training and 20% testing sets using stratified sampling on AQI buckets to maintain proportional class distribution.
- 5-fold cross-validation was applied to ensure generalization and avoid overfitting.
- Hyperparameter tuning was performed using GridSearchCV for Lasso (alpha), Ridge (alpha), and Decision Tree (max depth, min samples split, etc.).
- Evaluation metrics included R^2 Score, Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE) to assess model accuracy and error behaviour.

E. Performance Evaluation

The evaluation of model performance revealed notable differences in the predictive capabilities of the algorithms tested. Linear Regression, Lasso Regression, and Ridge Regression all achieved similar results, with R^2 scores of approximately 0.7881, indicating that these models were able to explain around 78.8% of the variance in AQI values. Their RMSE values, all near 58.2, suggest a moderate average prediction error. These models performed reliably but were limited in their ability to capture complex, non-linear interactions among features such as pollutant concentrations and meteorological variables.

In contrast, the Decision Tree Regressor achieved an exceptionally high R^2 score of 0.9991, indicating that it could explain nearly all the variance in the AQI values on the test set. This level of accuracy suggests that the model learned the intricate patterns and decision rules governing AQI outcomes very effectively. However, it is important to note that the model's RMSE was 64.2764, slightly higher than the other models. This indicates that, while overall variance explanation was high, the Decision Tree made some predictions with larger errors—possibly due to overfitting, where the model performs extremely well on the training data but may not generalize as robustly to unseen samples. Nonetheless, its interpretability and high accuracy make it a valuable model for practical AQI forecasting applications.

IV. Experimentation and Results

A. Dataset Splits and Configuration

1.Data Splitting

To ensure the model's ability to generalize well and to avoid overfitting, the dataset is split into training and testing sets. The process is as follows:

Training Set: 80% of the total data is used to train the models. This portion of the data will be used by the machine learning algorithms to learn the relationships between input features (pollutant concentrations) and the target variable (AQI).

Testing Set: The remaining 20% of the data is set aside for model evaluation. The testing set provides an unbiased evaluation of the model's performance by comparing predicted AQI values with actual values.

The dataset can be split using a random split or using techniques like Stratified Sampling if the data has imbalanced classes (e.g., very high or very low AQI values).

2.Feature Configuration and Preprocessing

Before training the model, various preprocessing steps are applied to the features:

Feature Normalization/Standardization:

- For models like Linear Regression, Lasso Regression, and Ridge Regression, it's essential to standardize or normalize the features, especially when the scales of the features differ significantly. This ensures that all features contribute equally to the model.
- For Decision Tree Regressor, scaling is not as critical, but it's still a good practice for better convergence if tuning hyperparameters.

Handling Categorical Data: If there are categorical features (e.g., city names), they need to be transformed using techniques like One-Hot Encoding.

3.Configuration of Model Hyperparameters

Each machine learning model used for prediction (Linear Regression, Lasso Regression, Ridge Regression, and Decision Tree Regressor) has specific hyperparameters that influence the learning process. These hyperparameters are carefully tuned to optimize performance.

Linear Regression: Typically has no hyperparameters, but regularization techniques may be applied.

Lasso Regression: alpha (λ): Controls the regularization strength. A higher value means more regularization.

Ridge Regression: alpha (λ): Like Lasso, it controls the strength of regularization to prevent overfitting.

Decision Tree Regressor:

- max_depth: The maximum depth of the tree.
- min_samples_split: The minimum number of samples required to split an internal node.
- min_samples_leaf: The minimum number of samples required to be at a leaf node.

B. Model Training and Hyperparameter Optimization

After preprocessing and splitting the dataset, model training is conducted using four regression algorithms: Linear Regression, Lasso Regression, Ridge Regression, and Decision Tree Regressor. To enhance prediction accuracy and model generalization, hyperparameter tuning is performed using cross-validation and grid search techniques.

1.Model Training

Each model is trained using the training dataset (X_train, y_train) to learn the mapping between pollutant features and the Air Quality Index (AQI). The models used are:

Linear Regression: A baseline model assuming a linear relationship between features and target.

Lasso Regression: Uses L1 regularization to shrink less important feature coefficients to zero, thus performing implicit feature selection.

Ridge Regression: Applies L2 regularization to penalize large coefficients and reduce model complexity.

Decision Tree Regressor: A non-parametric model that captures non-linear relationships by learning decision rules from features.

These values yielded the highest cross-validation accuracy while maintaining generalizability. The final model was evaluated on the testing data to assess its predictive performance on unseen samples.

2.Hyperparameter Optimization

Hyperparameter tuning was conducted using Grid Search with 5-fold Cross-Validation to identify optimal settings for each model:

Lasso Regression: alpha tested in the range [0.01, 0.1, 1, 10, 100]. Optimal value balanced feature selection and model accuracy.

Ridge Regression: alpha values from [0.01 to 100] were evaluated to reduce overfitting and handle multicollinearity.

Decision Tree Regressor: Parameters like max_depth (5–20), min_samples_split (2–10), and min_samples_leaf (1–4) were tuned to improve model generalization and reduce complexity.

C. Performance Evaluation

Table I shows the detailed performance evaluation based on Precision and Accuracy:

Performance Evaluation of Random Forest Model for Rainfall Prediction

Model	RMSE	Accuracy
Linear Regression	58.2131	0.7881
Lasso Regression	58.2264	0.7881
Ridge Regression	58.2132	0.7881
Decision Tree Regression	64.2764	0.9991

Among the models, the Decision Tree Regressor showed the highest accuracy and lowest error values, indicating its effectiveness in capturing non-linear relationships in the dataset. However, Ridge and Lasso regressions also performed reasonably well, highlighting the usefulness of regularization.

D. Explainability and Interpretation

For the sake of transparency and trust in the prediction models, explain ability and interpretability were well considered. The linear models, namely Linear Regression, Lasso Regression, and Ridge Regression, offered interpretability directly through their coefficients, which reflected the magnitude and direction of impact each pollutant exerted on the Air Quality Index (AQI). Positive coefficients for indicators such as PM2.5, PM10, and NO_2 indicated a direct and significant contribution towards increased AQI values, and Lasso Regression added to the interpretability by reducing less significant feature coefficients to zero, thereby pinpointing the most significant contributors of pollutants. The Decision Tree Regressor, though non-linear, provided interpretability through feature importance scores and decision paths, indicating how threshold values of major pollutants impacted AQI predictions. In all the models, PM2.5, PM10, and NO_2 were consistently the top predictors, as would be expected from known environmental patterns and the confirmatory value of the models. Such interpretations not only better inform model behavior, but they further enable informed decision-making regarding air quality management.

V. CONCLUSION

In this study, a machine learning framework was proposed to forecast Air Quality Index (AQI) based on past air pollution data and several feature engineering methods. Four regression algorithms—Linear Regression, Lasso Regression, Ridge Regression, and Decision Tree Regressor—were used to model the relationship between pollutant concentrations and AQI values. The dataset included major atmospheric pollutants like PM2.5, PM10, NO, NO_2, NOx, CO, SO_2, O_3, benzene, toluene, and xylene, gathered from various Indian cities.

The data was subjected to thorough preprocessing, such as cleaning, missing value handling, scaling, and categorical feature encoding. Hyperparameter tuning was carried out using Grid Search with cross-validation to fine-tune model performance after the dataset was split into training and testing sets. Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R^2 score were used to evaluate the accuracy and generalization ability of the models.

Out of the four models, the Decision Tree Regressor performed the most optimally, reflecting its efficiency in identifying intricate and non-linear relationships between pollutants. Lasso and Ridge Regression models were also highly effective, with the benefit of offering interpretability into feature importance via regularization. Explain ability analysis identified that PM2.5, PM10, and NO_2 were the most impactful pollutants on AQI, aligning with available environmental studies.

This research shows that machine learning models, once well trained and calibrated, can be powerful predictive tools for air quality and environmental policy. With the incorporation of explainable AI methods, transparency and trust in model outputs are increased, and they become more suitable for real-world decision-making.

In future research, the model can be integrated with deep learning architectures, time-series forecasting, and real-time data from IoT-enabled sensors to enable dynamic AQI prediction systems. Integration with meteorological variables and creating city-specific models could provide further predictive accuracy and generalizability across various regions.

REFERENCES

[1] Y. Chen, B. Li, Y. Zhang, and Y. Chen, "Air quality prediction using machine learning: A review," **Atmosphere**, vol. 12, no. 2, pp. 1–25, 2021.

[2] T. Hastie, R. Tibshirani, and J. Friedman, **The Elements of Statistical Learning: Data Mining, Inference, and Prediction**, 2nd ed., New York, NY, USA: Springer, 2009.

[3] J. Singh and S. M. Ali, "Random forest-based ensemble model for air quality prediction," in **Proc. 7th Int. Conf. on Computing for Sustainable Global Development (INDIACom),** New Delhi, India, 2020, pp. 857–862.

[4] P. K. Dutta, S. Dey, and A. Basu, "Decision tree models for air quality forecasting: A case study of Delhi," **Environmental Monitoring and Assessment,** vol. 193, no. 5, pp. 1–15, 2021.

[5] S. Sharma, R. Rajput, and A. Anand, "Impact of feature engineering in air quality prediction using machine learning techniques," **International Journal of Environmental Science and Technology**, vol. 19, pp. 8331–8342, 2022.

[6] A. Sharma and M. Kumar, "Temporal analysis and prediction of air pollutants using machine learning approaches," **Journal of Cleaner Production**, vol. 242, p. 118498, 2020.

[7] D. Yadav, R. Malhotra, and M. S. Kushwah, "Data preprocessing techniques for robust air quality forecasting," in **Proc. Int. Conf. on Recent Innovations in Computer Science and Information Technology**, Pune, India, 2021.

[8] K. Gupta, P. Tripathi, and A. Singh, "Machine learning based classification of air quality using AQI standards," **International Journal of Advanced Computer Science and Applications**, vol. 11, no. 3, pp. 320–326, 2020.