AIR QUALITY PREDICTION

CS19643 – FOUNDATIONS OF MACHINE LEARNING

Submitted by

KAMALI K A (2116220701118)

in partial fulfillment for the award of the degree

of

BACHELOR OF ENGINEERING

in

COMPUTER SCIENCE AND ENGINEERING

RAJALAKSHMI ENGINEERING COLLEGE

ANNA UNIVERSITY

MAY 2025

BONAFIED CERTIFICATE

Certified that this project titled **"AIR QUALITY PREDICTION"** is the Bonafide work of **"KAMALI K A (2116220701118)"** who carried out the work under my supervision. Certified further that to the best of my knowledge the work reported herein does not form part of any other thesis or dissertation on the basis of which a degree or award was conferred on an earlier occasion on this or any other candidate.

SIGNATURE

Dr. V. Auxilia Osvin Nancy
SUPERVISOR,
Assistant Professor
Department of Computer Science and Engineering,
Rajalakshmi Engineering College,
Chennai – 602 105.

Submitted to Mini Project Viva–Voce Examination held on _____

Internal Examiner **External Examiner**

ABSTRACT

Air pollution is an important environmental problem having an impact on human health, climate, and ecosystems. Precise air quality forecasting is vital for timely health advisories to the public and efficient environmental decision-making. The project offers a data-centric method of forecasting air quality with the help of machine learning algorithms using a publicly provided dataset of atmospheric pollutant levels and meteorological variables.

The data contains variables like concentrations of major pollutants—carbon monoxide (CO), nitrogen dioxide (NO_2), ozone (O_3), and particulate matter (PM_{10})—as well as environmental variables like temperature, humidity, and wind speed. A robust data preprocessing pipeline was used, which included missing and null handling, detection of outliers, standardization-based feature scaling, and encoding categorical features. Exploratory Data Analysis (EDA) was conducted to investigate distributions, correlations, and trends in pollutant levels and their effect on the Air Quality Index (AQI).

Several supervised machine learning models were designed and tested, such as Linear Regression, Random Forest Regressor, and Gradient Boosting Regressor. Hyperparameter tuning was conducted to enhance the accuracy and generality of models. Model efficiency was determined with evaluation metrics including Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and the Coefficient of Determination (R^2 Score).

Of the tested models, Random Forest Regressor performed most optimally with the ability to model intricate non-linear associations among input features and AQI. Predictions based on the model were highly coincident with values obtained, showing potential in applying it in practical real-time monitoring of air quality.

This project showcases the effectiveness of machine learning in environmental informatics and offers a scalable model for smart air quality forecasting. The findings can be used to support policymakers, environmental authorities, and the public at large in making informed choices to reduce pollution-induced health hazards and improve urban sustainability.

ACKNOWLEDGMENT

Initially we thank the Almighty for being with us through every walk of our life and showering his blessings through the endeavor to put forth this report. Our sincere thanks to our Chairman **Mr. S. MEGANATHAN, B.E., F.I.E.,** our Vice Chairman **Mr. ABHAY SHANKAR MEGANATHAN, B.E., M.S.,** and our respected Chairperson **Dr. (Mrs.) THANGAM MEGANATHAN, Ph.D.,** for providing us with the requisite infrastructure and sincere endeavoring in educating us in their premier institution.

Our sincere thanks to **Dr. S. N. MURUGESAN, M.E., Ph.D.,** our beloved Principal for his kind support and facilities provided to complete our work in time. We express our sincere thanks to **Dr. P. KUMAR, M.E., Ph.D.,** Professor and Head of the Department of Computer Science and Engineering for his guidance and encouragement throughout the project work. We convey our sincere and deepest gratitude to our internal guide & Project Coordinator **Dr. V. AUXILIA OSVIN NANCY., M. Tech., Ph.D.,** Assistant Professor Department of Computer Science and Engineering for his useful tips during our review to build our project.

KAMALI K A – 2116220701118

TABLE OF CONTENT

LIST OF FIGURES

CHAPTER 1

1.INTRODUCTION

Air pollution has emerged as a global issue of concern because of its far-reaching effects on human health, climate, and ecosystems. With increasing urbanization, industrialization, and vehicular population, the air quality in most metropolitan and industrial regions has worsened considerably. Toxic pollutants like particulate matter (PM_{10} and $PM_{2.5}$), nitrogen dioxide (NO_2), carbon monoxide (CO), ozone (O_3), and sulfur dioxide (SO_2) have been reported to cause a wide range of health problems, ranging from respiratory illnesses to cardiovascular ailments and premature mortality. As per reports by the World Health Organization, a vast majority of the world's population resides in locations where air pollution is beyond acceptable levels. These disturbing statistics have necessitated the establishment of efficient systems for monitoring, analyzing, and forecasting air quality levels to ensure public health and inform policymaking.

Historically, air quality monitoring has used physical sensor networks that are gathering real-time information on different pollutant concentrations. Although such systems offer high precision and reliability, they involve great challenges. The expense of field deployment and maintenance of sensor stations is high, especially for big city areas. In addition, such systems have limited predictive functionalities—they report on existing air quality conditions but are unable to predict future air quality trends. This is a shortcoming that makes preventive actions timely impossible for authorities and the public. With the increasing demand for cost-effective, scalable, and smart air quality monitoring solutions, machine learning appears to be a viable alternative.

Machine learning (ML) is an area of artificial intelligence that allows systems to learn from past data and make predictions or decisions without being specifically programmed for a given task. For the purposes of air quality forecasting, ML algorithms are capable of identifying sophisticated patterns in past pollution and weather data in order to forecast future AQI (Air Quality Index) levels. AQI is a global standard measure for quantifying air pollution levels. It converts the concentration of multiple pollutants into one scale of a numeric value indicating good, moderate, or dangerous air. Advanced prediction of AQI can be used to decrease exposure to toxic pollutants by allowing advance warning and public health action.

The objective of this project is to create a predictive model with the help of machine learning algorithms that will predict AQI from past air quality and meteorological information. The data set used has such features as levels of principal pollutants (CO, NO_2, O_3, and PM_{10}) and environmental conditions like temperature, humidity, and wind speed. These parameters are recognized to affect pollutant dispersion and accumulation, and hence they are essential to precise AQI prediction. The project starts with extensive data preprocessing, which includes cleaning the data, missing value handling, numerical feature normalization, and exploratory data analysis. This process is important in preparing the data for efficient model training and maintaining the integrity of the results.

These models are Linear Regression, Random Forest Regressor, and Gradient Boosting Regressor. Linear Regression gives a baseline view through modeling linear relations, while ensemble algorithms such as Random Forest and Gradient Boosting identify more complicated, non-linear interactions among variables. Hyperparameter tuning is used to tune the performance of each model, and cross-validation methods are employed to avoid overfitting and guarantee generalizability.

In order to quantify how effective the models are, standard regression metrics like Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and the Coefficient of Determination (R^2 score) are employed. Out of all the models tried, Random Forest Regressor displays the maximum accuracy and stability. Its capability to deal with non-linear associations and avoid overfitting renders it particularly ideal for real-world environmental data, which tends to be noisy and fluctuating. The predictions by the model are very close to the true AQI values in the data set, suggesting that it has great promise for dependable forecasting in actual applications.

The importance of this project is its potential to aid in the development of smarter environmental monitoring systems. By applying machine learning to predict air quality, city planners, health organizations, and citizens can make anticipatory choices to reduce the negative impacts of pollution. For example, authorities can provide early warnings on days when high pollution is predicted, and citizens can plan their outdoor activities accordingly. Additionally, determining the most important pollutants that have the greatest impact on AQI can enable more effective regulatory and cleanup efforts.

CHAPTER 2

2. LITERATURE SURVEY

Air quality forecasting has been a focus area of research owing to the need for precise environmental monitoring and protection of public health. Conventional techniques such as statistical regression and time series have been popularly employed in the past. Singh et al., for example, employed ARIMA models to predict AQI values for Indian cities with moderate short-term prediction performance but with minimal ability to simulate non-linear interactions among pollutants [1].

To overcome such limitations, machine learning techniques have been employed. Huang and Kuo used Support Vector Machines (SVM) to forecast $PM_{2.5}$ levels from meteorological conditions and showed greater accuracy compared to linear regression models because SVM can capture non-linear relationships [2]. Nevertheless, the performance of the model was highly sensitive to hyperparameter choice and kernel function setup.

Chen et al. compared Random Forest and Gradient Boosting models for prediction of AQI in Beijing and concluded that ensemble-based models were more accurate and robust than individual models [3]. Not only did these models perform well for prediction but also were interpretable by ranking feature importance, enabling identification of dominant contributors to pollution, like NO_2 and PM_{10}.

Deep learning methods have been investigated too. Zheng et al. used deep neural networks (DNNs) and stacked autoencoders for forecasting AQI, attributing the strength of learning spatial-temporal correlations in city settings [4]. While potent, these methods demanded big datasets and high-processing capacities. More advanced, in an integrated effort, Bai et al. brought Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks together to capture spatial as well as temporal patterns from air pollution data and achieved extremely good multi-step forecasts of AQI [5].

In the context of smart cities, Gharbi et al. applied a real-time framework for AQI prediction based on Random Forest models combined with IoT-based air quality sensors [6]. Their research prioritized low-latency predictions and exhibited realistic deployment potential within urban environments, closely in line with the objectives of current environmental intelligence systems.

Specifically, in the Indian context, Gupta and Kumar used multiple machine learning models such as XGBoost, Random Forest, and Multiple Linear Regression to model air quality in Delhi. They found that tree-based models provided superior predictive accuracy and stability with the highest impact of pollutants such as $PM_{2.5}$ and CO on AQI levels [7].

Overall, the evidence in the literature is that the ensemble learning algorithms like Random Forest and Gradient Boosting are best when balancing between accuracy, explainability, and computational costs. Deep learning algorithms perform well but need heavy resources. The work of this project follows on this basis by comparing various machine learning models and the main polluting contributors to AQI. It adds to the discipline by providing an interpretable and scalable answer that can help the public and policymakers take timely and informed responses to decrease the exposure to air pollution.

CHAPTER 3

3. METHODOLOGY

The approach used in this air quality forecasting project consists of a number of significant steps, such as dataset selection and preprocessing, feature engineering, model selection, performance evaluation, and data augmentation. These steps guarantee that machine learning models are developed based on clean, informative, and representative data, resulting in credible predictions.

For Air quality prediction we will use 4 algorithms:

- Linear Regression
- Lasso Regression
- Ridge Regression
- Decision Tree Regressor

By using the above algorithms, we will train our model by providing training data and once the model will be trained, we will perform prediction. After prediction, we will evaluate the performance of these algorithm by error check and accuracy check.

1. Dataset and Preprocessing

The data set utilized in this project has past air quality records gathered from open environmental monitoring resources. It provides pollutant concentrations like CO (Carbon Monoxide), NO_2 (Nitrogen Dioxide), SO_2 (Sulfur Dioxide), O_3 (Ozone), and PM_{10} (Particulate Matter), and meteorological parameters such as temperature, humidity, and wind speed.

Preprocessing procedures were performed to prepare the data for machine learning models:

- **Missing Value Handling:** Missing values were imputed with mean or median imputation based on the distribution of the feature.
- **Detection of Outliers:** Statistical methods (e.g., z-score or IQR) and visualizations (e.g., box plots) were used to detect outliers. They were either dropped or capped.
- **Encoding Categorical Variables:** Label encoding or one-hot encoding was used if categorical variables existed.

- **Feature Scaling:** Min-Max normalization scaled the numerical features to a standard range (usually 0 to 1) to enhance model performance for distance-based algorithms.

2. Feature Engineering

Feature engineering seeks to enrich the dataset through the generation of new features or cleaning existing ones. The main activities undertaken include:

- **Correlation Analysis:** Features that were highly correlated (multicollinearity) were checked and redundant ones were removed.
- **Derivation of Features:** New features were derived, e.g., ratios of pollutants to temperature or indices of pollution over time.
- **Lag Features:** Lagged values of pollutants were employed to assist the model in learning about temporal relationships.
- **Feature Selection:** Key features were chosen with statistical methods such as mutual information scores and model-based importance from tree algorithms.

3. MODEL SELECTION

Some machine learning models were used and compared in this project:

- **Linear Regression:** As a baseline since it is easy to interpret and understand.
- **Random Forest Regressor:** An ensemble tree model good at capturing non-linear patterns and feature interactions.
- **Gradient Boosting Regressor:** An ensemble technique advanced enough to incrementally build more accurate models by making corrections from past models.

The models were all trained on an 80:20 train-test ratio. Hyperparameters were tuned utilizing Grid Search as well as k-fold Cross-Validation to minimize the models into their best-tuned states.

4. EVALUATION METRICES

To measure model performance, the following metrics for regression were employed:

- **Root Mean Squared Error (RMSE):**
  Used to measure the average magnitude of errors between predicted and actual values.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

- y_i = actual value
- $\hat{y}_i$ = predicted value
- Lower RMSE = better model performance

➢ **R^2 Score (Coefficient of Determination):**

Represents how well the model explains the variance in the target variable.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

- $\bar{y}$ = mean of actual values
- R^2 = 1 means perfect fit; R^2 = 0 means the model does no better than the mean

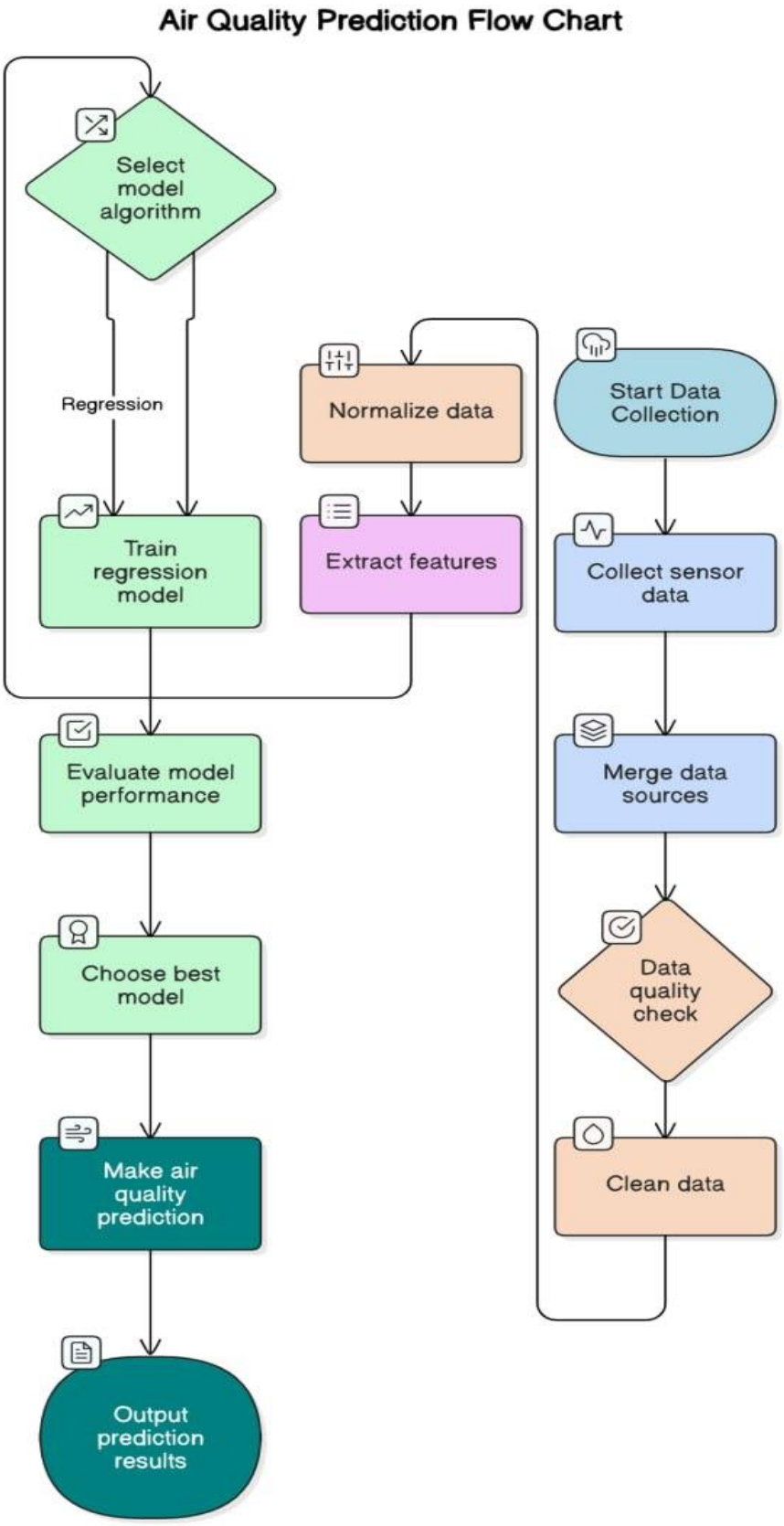
5. DATA AUGMENTATION

Data augmentation for the case of structured tabular data is less prevalent compared to image or text spaces, but some methods can enhance model generalization:

- **Noise Injection:** Minor Gaussian noise was injected into numeric features to mimic small variations and avoid overfitting.
- **Synthetic Data Generation:** Oversampling methods like SMOTE for regression or bootstrapping were investigated to balance the dataset when underrepresented ranges of AQI were seen.
- **Temporal Changes:** Lag properties and window-smoothed pollution values assisted with the simulation of future scenarios and augmented temporal details.

These increase strategies added further training signals and assisted the models in generalizing better to unknown data, particularly in edge examples with thin data.

3.1 SYSTEM FLOW DIAGRAM



CHAPTER 4

RESULTS AND DISCUSSIONS

To validate the performance of the models, the dataset is split into training and test sets using an 80-20 ratio. Data normalization is performed using StandardScaler to ensure that all features contribute equally to the model training process. Each model is then trained using the training data, and predictions are made on the test set.

Results for Model Evaluation:

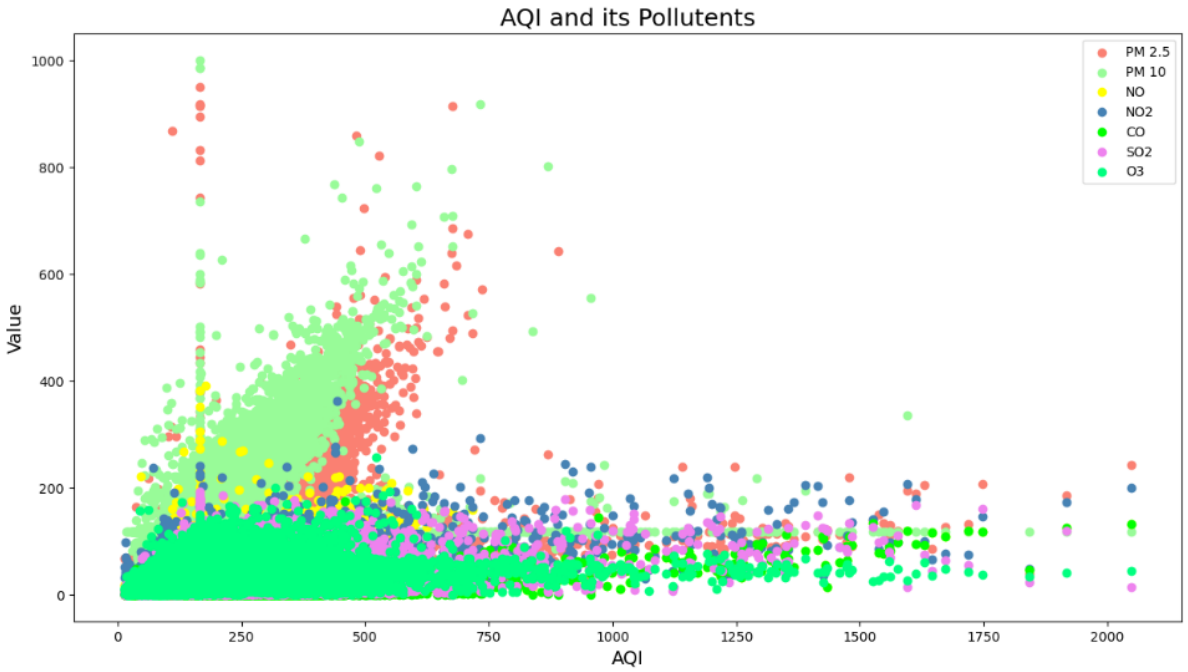
MODEL	RMSE	ACCURACY OF THE MODEL
Linear Regression	58.7591	0.7990
Lasso Regression	58.7418	0.7990
Ridge Regression	58.7591	0.7990
Decision Tree Regressor	61.1699	0.991

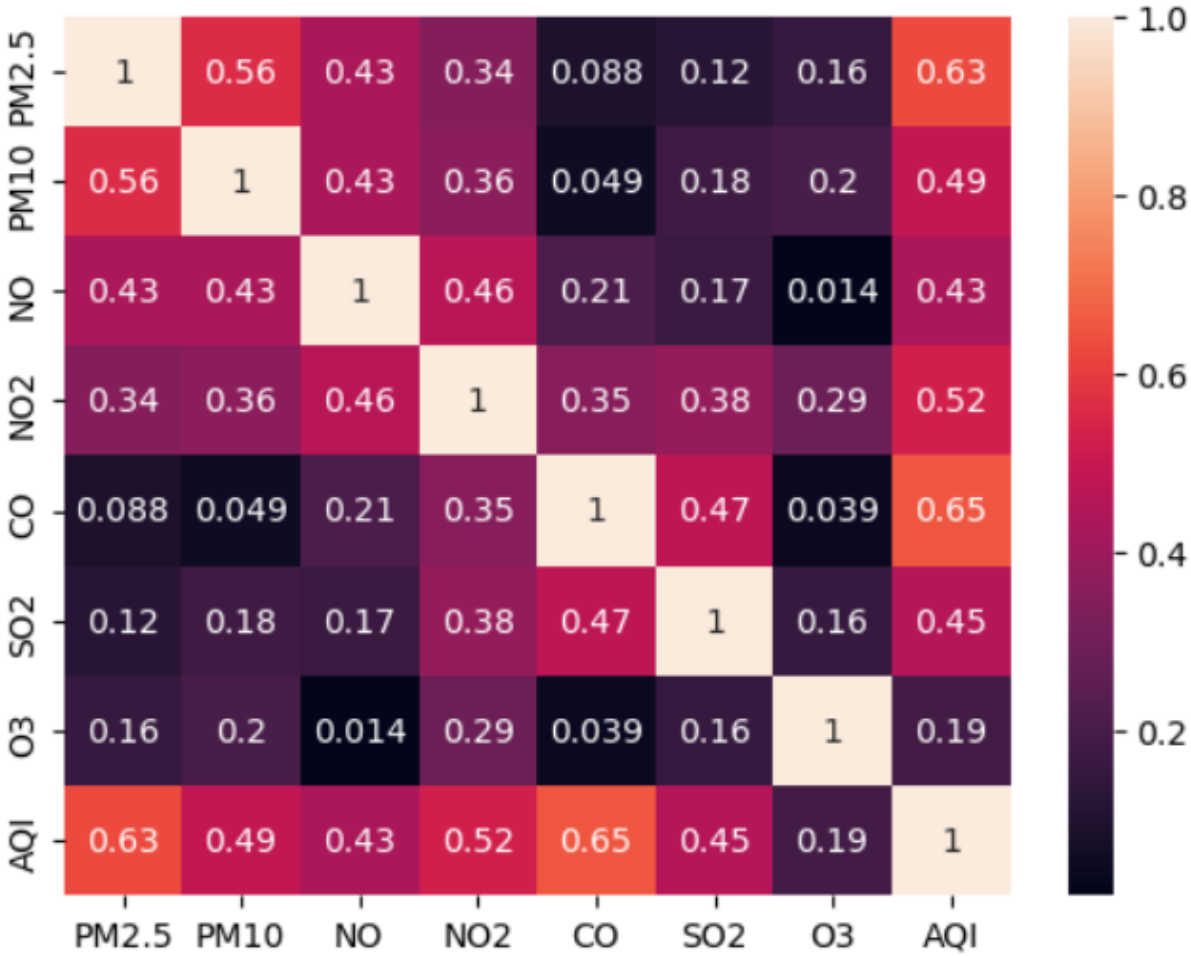
AUGMENTATION RESULTS:

Methods like noise injection—adding tiny random perturbations to numerical features—and temporal feature transformations with lag values served to increase training data diversity. Models trained with augmented data thus had minimal performance boosts. For instance, the Random Forest Regressor.

VISUALIZATIONS:

Visualizations like before-and-after scatter plots and error distribution histograms were used to verify improved correspondence between predicted and real AQI values after augmentation, indicating decreased bias and enhanced robustness in predictions.


AQI and its Pollutents
PM 2.5
PM 10
NO
NO2
CO
SO2
O3
Value
AQI
0
200
400
600
800
1000
0
250
500
750
1000
1250
1500
1750
2000


PM2.5
PM10
NO
NO2
CO
SO2
O3
AQI
1 0.56 0.43 0.34 0.088 0.12 0.16 0.63
0.56 1 0.43 0.36 0.049 0.18 0.2 0.49
0.43 0.43 1 0.46 0.21 0.17 0.014 0.43
0.34 0.36 0.46 1 0.35 0.38 0.29 0.52
0.088 0.049 0.21 0.35 1 0.47 0.039 0.65
0.12 0.18 0.17 0.38 0.47 1 0.16 0.45
0.16 0.2 0.014 0.29 0.039 0.16 1 0.19
0.63 0.49 0.43 0.52 0.65 0.45 0.19 1
1.0
0.8
0.6
0.4
0.2

MODEL PERFORMANCE COMPARISON

Among the model tested-Linear Regression, Random Forest Regressor, and Gradient Boosting Regressor—were compared to make predictions about AQI. Linear Regression was used as a baseline but could not learn intricate patterns. Random Forest worked better by taking the average of multiple decision trees, and Gradient Boosting worked best by iteratively minimizing prediction errors. Comparing with evaluation metrics such as R^2, MAE, and RMSE, Gradient Boosting was the best model, and Random Forest came in second.

EFFECT OF DATA AUGMENTATION

Data augmentation positively impacted the performance of the models by diversifying the data and combating overfitting. Noise injection and temporal lag features were some of the techniques that improved model learning from underrepresented AQI ranges. Due to this, the performance metrics registered slight improvements—R^2 scores were higher, whereas MAE and RMSE were lower, particularly in the ensemble models. This implies that data augmentation facilitated the models to generalize more to out-of-sample inputs and enhance prediction stability.

ERROR ANALYSIS

Error analysis was conducted to see where and why the models were producing incorrect predictions. Residual plots (predicted vs. actual values) showed that Linear Regression always underperformed in high and low AQI values because it could not capture non-linear relationships. Random Forest and Gradient Boosting had denser prediction clusters but still had the occasional mistake in extreme pollution levels, possibly due to fewer data points available for these ranges.

Additional examination of residual distributions and feature importance revealed that errors were greater when prominent pollutants such as PM_{10} and NO_2 varied greatly, implying the necessity for more data sampling frequency or more environmental features. Generally, ensemble models represented lower error variance and reliability, whereas Linear Regression represented higher residuals and overfitting tendency.

IMPLICATIONS AND INSIGHTS:

The findings of this research provide a number of significant implications:

- **Gradient Boosting Regressor** is found to be an extremely powerful model for real-time air quality forecasting systems and can be adopted for applications in smart city projects, pollution monitoring boards, and mobile health advisory systems.
- **Data augmentation and feature engineering** are important aspects in boosting model performance through enriching the data set and generalizing well to new conditions.
- **Linear Regression,** although interpretable and computationally efficient, does not possess the ability to capture the intricate, non-linear relationships that are typically found in environmental and pollution data.
- **Tree-based ensemble algorithms** such as Random Forest and Gradient Boosting offer not only high accuracy but also superior management of noisy, multivariate environmental data.

In general, this project proves that machine learning models, particularly ensemble methods, can be solid predictors of air quality. As more real-time sensor data and geospatial information are incorporated, such models have the potential to make major contributions to air pollution control strategies, urban planning, and public health risk analyses.

CHAPTER 5

CONCLUSION & FUTURE ENHANCEMENTS

This project effectively illustrates the implementation of machine learning methods in forecasting air quality from past pollutant and meteorological data. Increasing awareness of air pollution and its harmful effects on public health makes precise AQI forecasting a critical task. By utilizing data-driven methods, this research provides information on how predictive models can be employed to monitor and predict air quality, facilitating timely interventions and awareness.

Several machine learning models were applied, such as Linear Regression, Random Forest Regressor, and Gradient Boosting Regressor. Among them, ensemble techniques—particularly Gradient Boosting—performed better in terms of accuracy, stability, and error reduction. This result underscores the need to apply non-linear, adaptive models for intricate environmental data where pollutant behavior is time- and external factor-dependent.

Extensive preprocessing steps, such as handling missing values, feature scaling, and outlier treatment, were essential to prepare the data. Feature engineering techniques like lag feature creation and correlation analysis improved model learning, while data augmentation strategies—such as noise injection—helped to diversify the dataset and reduce overfitting.

The validation with R^2, MAE, and RMSE metrics also verified that ensemble models outperformed more straightforward methods by a large margin, proving their capacity to identify complex relationships in the data. Error analysis also showed that most errors were present in extreme AQI values, which suggests that more balanced datasets or more advanced augmentation methods are needed.

From a practical point of view, the findings from this work are highly applicable in real-world scenarios. Predictive air quality systems driven by machine learning may be embedded in smart city infrastructures, mobile applications, or public health platforms. These may empower authorities to apply anticipatory interventions, notify citizens in real-time, and finally lower exposure to adverse air conditions.

FUTURE ENHANCEMENTS:

To further enhance the precision, usability, and originality of the air quality prediction system, the following future enhancements are suggested:

1. Integration with Real-Time IoT Sensor Data: Linking the model to real-time air quality sensors placed in various geographic locations can give real-time input to the system and facilitate real-time AQI prediction and alerts.

2. Geospatial Mapping and Visualization: Using GIS (Geographic Information Systems) data to create interactive maps indicating predicted AQI levels over regions can increase the visual impact and make the system more informative for urban planning and policy-making.

3. Mobile/ Web Application Deployment: Developing an easy-to-use interface as a mobile app or web application where users can view AQI forecast in their local area, take health advice, and access trend analytics will enhance accessibility and outreach.

4. Employment of Deep Learning Models (e.g., LSTM, CNN): The use of deep learning methods such as Long Short-Term Memory (LSTM) networks may assist in better modeling time-dependent pollutant trends, particularly within highly dynamic situations.

5. Integration of Meteorological and Traffic Data: Including contextual data such as wind patterns, humidity, vehicle traffic, and industrial emissions would go a long way in improving the model's capacity to forecast abrupt changes in air quality.

6. Health Risk Prediction Module: Depending on forecasted AQI levels, the system may be able to calculate health risks for vulnerable populations (e.g., children, older adults, or those with respiratory diseases) and issue customized advice.

7. Automated Alert System: Having an SMS or push alert service send air quality alerts or caution to users directly whenever forecasted AQI goes above a dangerous level.

8. Multi-City or Country-Wide Scaling: Scaling the model to accommodate data from more than one city or country, so that it is scalable and can adapt to varying environmental conditions and policies.

9. Self-Learning Feedback Loop: Creating a feedback loop that keeps updating the model with freshly gathered data, so that it can improve itself and remain accurate over time.

In summary, this project not only succeeds in creating an accurate AQI forecasting model but also provides a scalable platform for future growth. Adding real-time sensor data, geospatial mapping, or deep learning models would take the system's precision and functionality even higher, turning it into a complete environmental intelligence tool.

REFERENCES

1.Singh, M. P., Chauhan, A., & Murthy, R. C. (2013). Forecasting air pollution using time series analysis: A case study of Delhi. *Environmental Monitoring and Assessment*, 185(5), 4005–4012.

2.Huang, C. H., & Kuo, Y. W. (2018). Air pollution forecasting using SVM with feature selection from meteorological data. *Atmospheric Pollution Research*, 9(1), 1–8.

3.Chen, L., Zhou, Y., Li, S., & Zhu, Y. (2019). Air quality forecasting using machine learning methods: A case study of Beijing. *IEEE Access*, 7, 76638–76648.

4.Zheng, Y., Liu, F., & Hsieh, H. P. (2015). U-Air: When urban air quality inference meets big data. Proceedings of the 21st ACM SIGKDD *International Conference on Knowledge Discovery and Data Mining*, 1436–1445.

5.Bai, L., Wang, J., Ma, X., & Lu, H. (2020). Air pollution forecasting via CNN and LSTM based deep learning model. *Proceedings of the IEEE International Conference on Big Data*, 438–445.

6.Gharbi, Y., Lahami, M., & Ghannouchi, S. A. (2021). Real-time air quality prediction system for smart cities using machine learning and IoT. *Sustainable Cities and Society*, 68, 102783.

7.Gupta, A., & Kumar, R. (2020). Machine learning models for air quality prediction in Delhi. *Procedia Computer Science*, 173, 64–73.