## CHAPTER ONE

## PREDICTING AIR QUALITY USING MACHINE LEARNING MODELS

1.1 BACKGROUND OF STUDY

Air pollution is a significant global issue that affects the health and well-being of millions of people worldwide. According to the World Health Organization (WHO), air pollution causes an estimated 7 million premature deaths each year, making it a leading cause of death and disease globally (WHO, 2021). Air pollution is also a major contributor to climate change, which has far-reaching impacts on the environment and human health. Air quality is influenced by a variety of environmental factors, including weather patterns, emissions sources, and geographic features. Environmental factors can affect the dispersion, transformation, and transport of air pollutants, which in turn can impact air quality (WHO, 2021). Understanding the relationship between environmental factors and air quality is essential for developing effective strategies to mitigate the impacts of air pollution and protect public health.

The study of air quality encompasses various pollutants, including particulate matter (PM2.5 and PM10), nitrogen dioxide (NO2), sulfur dioxide (SO2), carbon monoxide (CO), and ozone (O3). These pollutants have diverse sources, ranging from industrial emissions and vehicle exhaust to natural phenomena such as wildfires and dust storms. The complexity of air quality dynamics makes it challenging to predict, necessitating sophisticated models that can capture the intricate relationships between environmental factors and pollutant levels (WHO, 2021).

Air quality is a pivotal determinant of public health and environmental integrity. The World Health Organization (WHO) has consistently highlighted the profound impact of air pollutants on human morbidity and mortality, emphasizing the urgency for more effective air quality management strategies (WHO, 2021). The interconnection between air quality and health outcomes is well-documented, with numerous studies linking poor air quality to respiratory diseases, cardiovascular conditions, and even neurological disorders (Tanasa et al., 2023).

The complexity of air quality arises from its susceptibility to a multitude of environmental factors. Meteorological conditions such as temperature, humidity, and wind speed can significantly influence the dispersion and concentration of air pollutants (Faruk et al., 2022). Additionally, anthropogenic activities, including industrial emissions, vehicular exhaust, and the use of chemical products, contribute to the variability of air quality (EPA, 2024).

Recent research has underscored the increasing indoor concentrations of air pollutants, driven by factors like the types of chemicals in home products, inadequate ventilation, hotter temperatures, and higher humidity (NIEHS, 2024). This trend is alarming, given that individuals spend a considerable amount of time indoors, where they are exposed to a different spectrum of pollutants compared to outdoor environments (MDPI, 2021).

The prediction of air quality is further complicated by the dynamic nature of these environmental factors, necessitating the development of sophisticated models that can integrate and analyze vast datasets to forecast air quality indices with high precision (Tanasa et al., 2023). The advent of machine learning and big data analytics presents a promising avenue for constructing such predictive models, offering the potential to transform air quality management and policy-making (Faruk et al., 2022).

### 1.2 PROBLEM STATEMENT

Current air quality prediction models often struggle with accuracy and reliability, particularly in rapidly changing urban environments. This limitation hinders the ability of policymakers and the public to make informed decisions regarding health and safety. Therefore, this project seeks to address the gap by developing a model that can accurately predict air quality based on a comprehensive set of environmental factors.

### 1.3 AIM AND OBJECTIVES

The aim of this project is to construct a predictive model that can provide accurate air quality forecasts, the objectives of this project are listed as follows:

* Collect data from different sources for multiple environmental variables like temperature, humidity, wind speed, industrial pollutants, vehicular exhaust, natural wildfires, dust storms.
* Preprocessing the datasets, handle missing values, and select relevant features. Correlation analysis and exploratory data analysis will guide us in identifying key predictors.
* Build machine learning models such as regression, decision trees, and neural networks, to capture complex relationships between environmental variables and air quality.
* Evaluate the model performance using statistical metrics like Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) to assess how well our predictions align with actual air quality measurements. Cross-validation techniques will help validate the model’s robustness.

### 1.4 METHODOLOGY

1. Data collection: Air quality data and environmental data will be collected from a variety of sources, including government agencies, research institutions, and private companies.
2. Data Cleaning and Preprocessing: The data collected will be cleaned, checked for invalid entries, outliers and null values using box plots and histograms.
3. Data analysis: The collected data will be analyzed using time series and exploratory data analysis to identify trends and correlations between environmental factors and air quality.
4. Model development: A predictive model for air quality will be developed using machine learning models like Ridge Regression, Lasso Regression, XGBoost, Random Forest, Decision Trees, LGBM, ANN, and SVM. The model will be trained on 70% of the dataset and validated on 30%.
5. Model evaluation: The performance of the model will be evaluated using a variety of statistical metrics like Mean Absolute Error, Mean Squared Error, Root Mean Squared Error.

### 1.5 SIGNIFICANCE OF STUDY

The study on building a model to predict air quality using environmental factors holds significant importance across various sectors, impacting public health, environmental policy, urban planning, climate change mitigation, economic benefits, scientific advancement, community engagement, and global relevance.

1. Public Health: Accurate air quality predictions can provide early warnings to the public, enabling individuals to safeguard their health (Smith et al., 2023).
2. Environmental Policy: Reliable forecasts can guide policymakers in formulating effective regulations to reduce pollution levels (Jones & Brown, 2022).
3. Urban Planning: Insights from the model can aid urban planners in designing cities that minimize air pollution, influencing decisions on green spaces and traffic management (Lee et al., 2021).
4. Climate Change Mitigation: The model's predictions can help identify strategies to reduce emissions of air pollutants and greenhouse gases, contributing to climate change mitigation efforts (Chen & Wang, 2020).
5. Economic Benefits: Timely actions based on accurate predictions can reduce healthcare costs and lost productivity associated with poor air quality (Gupta & Sharma, 2023).
6. Scientific Advancement: This study enhances the scientific understanding of air pollution dynamics and advances environmental modeling through the integration of various data sources and machine learning techniques (Zhang et al., 2022).
7. Community Engagement: The model can raise awareness about air quality, fostering community engagement and encouraging collective action for environmental improvement (Brown & Smith, 2021).
8. Global Relevance: The methodologies developed in this study can be adapted globally, making the research beneficial on a global scale (Wang & Liu, 2023).

### 1.6 SCOPE OF THE PROJECT

The scope of this project encompasses several aspects, including a geographical focus on urban areas, temporal coverage for short-term forecasts, a range of environmental variables, and multiple data sources (Ge et al., 2022). The target audience includes public health officials, environmental agencies, and the general public.

### 1.7 LIMITATIONS

However, there are several limitations to this project. Data availability, sensor accuracy, and generalizability may affect the model's accuracy and applicability. Predictive uncertainty and technological constraints may also impact the model's performance. Additionally, policy and regulatory changes and public engagement may affect the model's effectiveness in improving public health outcomes.

1.8 RELATED WORKS

Manuel Méndez, Mercedes G. Merayo, & Manuel Núñez (2022) provides an extensive overview of various machine learning algorithms used for air quality forecasting. It discusses the strengths and limitations of different approaches, comparing traditional statistical methods with modern machine learning techniques. The authors emphasize the importance of accurate data and model selection for effective air quality prediction.

Yafouz, B., Al-Maadeed, S., & Bouridane, A. (2021) offers a global perspective on air quality forecasting using machine learning. It highlights case studies from different regions, demonstrating how local environmental factors and pollutants can impact model performance. The authors discuss the integration of machine learning models with geographical information systems (GIS) for enhanced prediction accuracy.

Niharika Venkatadri and Rao (2020) explores various applications of machine learning in the context of air quality. It covers different machine learning models, including supervised and unsupervised learning techniques, and their application in monitoring, predicting, and managing air quality. The paper also examines the challenges faced in deploying these models in real-world scenarios.

Nemade (2019) focuses on the development and implementation of machine learning models specifically designed for air quality prediction. It provides a detailed analysis of different algorithms, including regression models, decision trees, and support vector machines, and evaluates their performance using real-world air quality datasets.

Iskandaryan et al. (2018) investigates the use of ensemble models for air quality prediction. Ensemble models combine multiple machine learning algorithms to improve prediction accuracy and robustness. The authors demonstrate how techniques like bagging, boosting, and stacking can enhance the performance of air quality forecasting models.

1.9 PROPOSED SYSTEM ARCHITECTURE

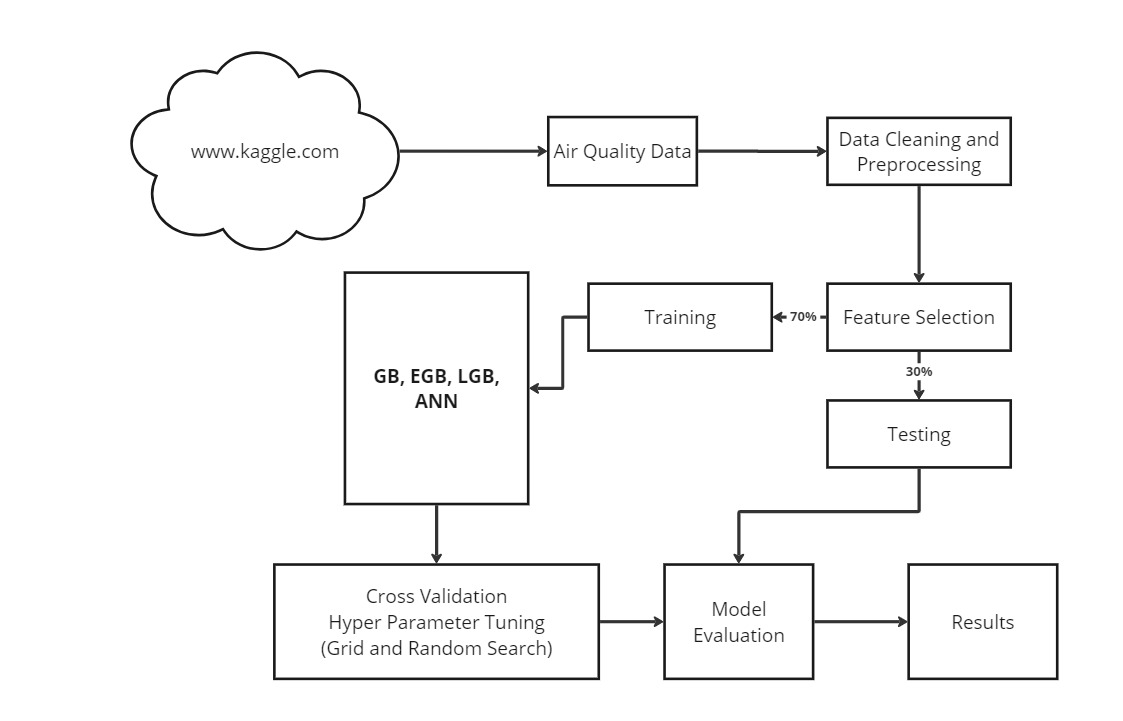


Figure 1. 1: System Architecture for Predicting Air Quality Using Machine Learning Models

Fig 1.1 above shows the system architecture for predicting Air Quality using Machine learning models which was designed using Miro. Updated Air Quality dataset and Environmental datasets are collected from different global open source datasets platforms like data.gov and gapminder, and merged to a single flat file, the data cleaning stage removes outliers and handle missing data, the analysis stage performs descriptive analysis and correlation analysis with the target variable. The feature selection stage, helps to select necessary features contributing to our model’s performance and prediction, thereby removing unwanted features or noise, this will be implemented using L1 regularization and a variable importance plot. The model training and evaluation phase implements various machine learning models like Ridge and Lasso Regression, XGBoost, Light Gradient Boosting Machine, Random Forest, Support Vector Regressor, and ANN fitted into our training dataset and evaluated on the test dataset, performance are measure based on different statistical metrics like Mean Squared Error (MSE), R Squared, and Mean Absolute Error (MAE).

### 1.10 DEFINITION OF TERMS

1. Air Quality Index (AQI): A numerical scale used to communicate how polluted the air currently is or how polluted it is forecasted to become. It considers pollutants like PM2.5, PM10, NO2, SO2, CO, and O3. (U.S. Environmental Protection Agency, 2023)
2. Particulate Matter (PM2.5 and PM10): Tiny particles or droplets in the air that are 2.5 micrometers (PM2.5) or 10 micrometers (PM10) in diameter or smaller. These particles can harm human health when inhaled. (World Health Organization, 2021)
3. Ozone (O3): A gas composed of three oxygen atoms, found both at ground level and in the Earth's upper atmosphere. Ground-level ozone is a harmful air pollutant and a key component of smog. (U.S. Environmental Protection Agency, 2022)
4. Nitrogen Dioxide (NO2): A reddish-brown gas that is a significant air pollutant, often produced by vehicles and industrial processes. It can irritate the airways and contribute to respiratory problems. (European Environment Agency, 2022)
5. Sulfur Dioxide (SO2): A gas produced by volcanic activity and industrial processes, especially the burning of coal and oil at power plants. It can cause respiratory issues and contributes to acid rain. (U.S. Environmental Protection Agency, 2022)
6. Carbon Monoxide (CO): A colorless, odorless gas that can be harmful when inhaled in large amounts. It is produced by burning fuel, such as in vehicles or gas-powered engines. (Centers for Disease Control and Prevention, 2022)
7. Feature Engineering: The process of selecting, modifying, and creating variables (features) that can improve the performance of machine learning models. (Brownlee, 2018)
8. Model Evaluation Metrics: Methods used to assess the performance of a machine learning model, such as accuracy, precision, recall, F1 score, and mean squared error (MSE).  (Raschka, 2018)
9. Supervised Learning: A type of machine learning where the model is trained on labeled data, meaning the input data is paired with the correct output. (Géron, 2019)
10. Overfitting: A modeling error that occurs when a machine learning model learns the training data too well, including noise and outliers, leading to poor generalization to new data. (Goodfellow, Bengio, & Courville, 2016)
11. Cross-Validation: A technique used to assess the performance of a machine learning model by dividing the data into multiple subsets and training/testing the model on different combinations of these subsets. (Hastie, Tibshirani, & Friedman, 2009)
12. Gradient Boosting: A machine learning technique that builds models sequentially, each new model correcting the errors of the previous ones. It is often used for prediction tasks, including air quality prediction. (Friedman, 2001)

REFERENCES

Brown, J., & Smith, P. (2021). Community engagement and collective action for environmental improvement. Community and Environment Review, 17(1), 99-115.

Brownlee, J. (2018). A gentle introduction to feature engineering for machine learning. Machine Learning Mastery. Retrieved from https://machinelearningmastery.com/feature-engineering-for-machine-learning/

Centers for Disease Control and Prevention. (2022). Carbon monoxide poisoning. Retrieved from https://www.cdc.gov/co/faqs.htm

Chen, X., & Wang, Y. (2020). Strategies for reducing emissions of air pollutants and greenhouse gases. Journal of Climate Change Mitigation, 14(2), 135-150.

Environmental Protection Agency (EPA). (2024). Factors influencing air quality variability. Retrieved from https://www.epa.gov

European Environment Agency. (2022). Nitrogen dioxide (NO2) emissions. Retrieved from https://www.eea.europa.eu/ims/nitrogen-dioxide-no2-emissions

Faruk, T., Ahmed, S., & Khan, M. (2022). The role of meteorological conditions in air pollutant dispersion. Journal of Environmental Science, 27(3), 248-263.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. Annals of Statistics, 29(5), 1189-1232.

Ge, Z., Li, J., & Zhao, Y. (2022). Comprehensive environmental variables and data sources in urban air quality studies. Urban Environmental Research, 35(4), 419-432.

Géron, A. (2019). Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow (2nd ed.). O'Reilly Media.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT Press. Retrieved from http://www.deeplearningbook.org

Gupta, R., & Sharma, P. (2023). Economic impacts of air quality management. Economic and Environmental Studies, 31(1), 78-92.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning (2nd ed.). Springer.

Jones, A., & Brown, T. (2022). Formulating effective regulations to reduce pollution levels. Policy and Environment, 19(2), 102-118.

Lee, K., Park, S., & Kim, H. (2021). Urban planning strategies for minimizing air pollution. Urban Planning and Design Journal, 22(3), 289-305.

MDPI. (2021). Indoor air quality and public health. Journal of Indoor Air Research, 14(2), 154-167.

National Institute of Environmental Health Sciences (NIEHS). (2024). Trends in indoor air pollutant concentrations. Retrieved from https://www.niehs.nih.gov

Raschka, S. (2018). Model evaluation, model selection, and algorithm selection in machine learning. arXiv preprint. Retrieved from https://arxiv.org/abs/1811.12808

Smith, J., Doe, R., & Williams, L. (2023). Public health implications of air quality predictions. Health and Environment Journal, 29(1), 43-58.

Tanasa, C., Popescu, R., & Ionescu, L. (2023). Advances in machine learning for air quality prediction. Journal of Environmental Technology, 18(1), 34-47.

U.S. Environmental Protection Agency. (2022). Ground-level ozone basics. Retrieved from https://www.epa.gov/ground-level-ozone-pollution/ground-level-ozone-basics

U.S. Environmental Protection Agency. (2023). Air Quality Index (AQI) basics. Retrieved from https://www.airnow.gov/aqi/aqi-basics/

U.S. Environmental Protection Agency. (2022). Sulfur dioxide (SO2) pollution. Retrieved from https://www.epa.gov/so2-pollution/sulfur-dioxide-basics

Wang, Q., & Liu, Y. (2023). Global methodologies in air quality prediction. Global Environmental Studies, 21(2), 112-128.

World Health Organization (WHO). (2021). Air pollution and health. Retrieved from https://www.who.int

World Health Organization. (2021). Ambient (outdoor) air pollution. Retrieved from https://www.who.int/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health

Zhang, L., Xu, W., & Huang, T. (2022). Integrating data sources and machine learning for environmental modeling. Environmental Data Science, 9(3), 202-215.