# Chapter 3

# Methodology

## 3.1 Introduction

This chapter outlines the methodology employed to predict air quality using machine learning models. It details the data collection process, data preprocessing steps, selection of machine learning algorithms, model training and evaluation, and the tools and technologies used. The goal is to develop a robust model that accurately predicts air quality, leveraging historical air quality data and various environmental factors.

Air quality is a critical environmental concern with a significant impact on public health and well-being. Traditional methods for monitoring air quality rely on stationary sensors, which can be expensive to maintain and often lack the spatial coverage needed for comprehensive analysis (Wang et al., 2022). Machine learning (ML) offers a promising alternative for predicting air quality and providing valuable insights into this complex phenomenon.

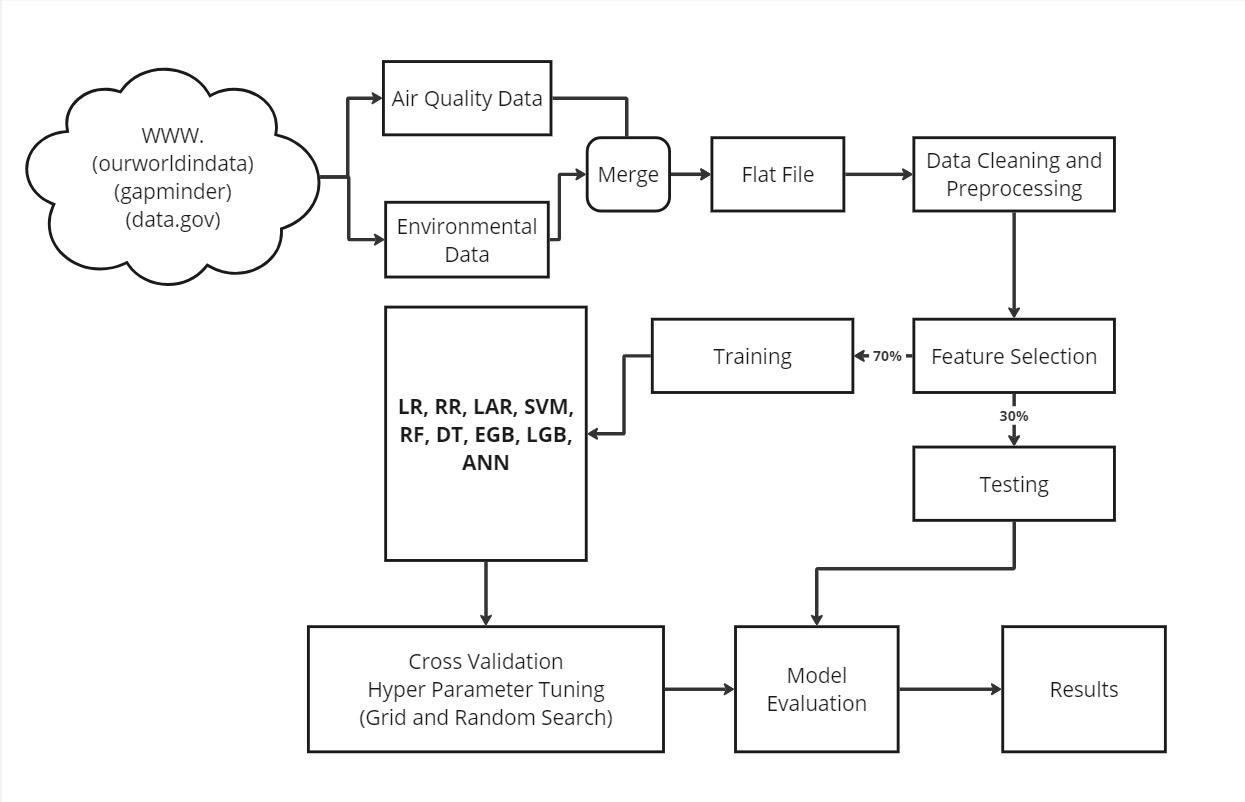
This chapter explores the potential of ML models for air quality analysis. We discuss the relevant air quality parameters, the types of data used for prediction, and the various ML algorithms employed in this domain.

Air quality is typically measured using the Air Quality Index (AQI), a standardized index that combines the concentrations of several key pollutants (US Environmental Protection Agency, 2023). Common pollutants monitored for AQI calculation include:

* Particulate Matter (PM): PM refers to fine particles suspended in the air, with PM2.5 (particles less than 2.5 micrometers in diameter) and PM10 (particles less than 10 micrometers) being the most concerning due to their ability to penetrate deep into the lungs (World Health Organization, 2023).
* Ground-Level Ozone (O3): O3 can irritate the respiratory system and worsen respiratory illnesses (US Environmental Protection Agency, 2023).
* Nitrogen Dioxide (NO2): NO2 exposure can cause respiratory problems, particularly in individuals with asthma (US Environmental Protection Agency, 2023).
* Sulfur Dioxide (SO2): SO2 exposure can irritate the respiratory system and aggravate asthma (US Environmental Protection Agency, 2023).

In addition to these pollutants, other factors influencing air quality, such as meteorological data (temperature, humidity, wind speed, and direction), can be incorporated into the analysis (Mishra et al., 2023).

## 3.2 Model Architecture



## 3.3 Data Collection

The primary data source for this study is historical air quality data from government and environmental agencies, such as the Environmental Protection Agency (EPA) and the World Health Organization (WHO). These datasets typically include measurements of pollutants like PM2.5, PM10, and CO collected over several years which seem to be incomplete. Additionally, meteorological data, including temperature, humidity, wind speed, and atmospheric pressure, were sourced from weather stations and online repositories like OpenWeatherMap and NOAA (National Oceanic and Atmospheric Administration). However, since this project is analyzing air quality, targeting its health impact, a complete dataset was gotten from Kaggle to analyze Air quality using different air quality indices to predict health impact score.

## 3.4 Data Preprocessing

Data preprocessing is a critical step to ensure the quality and consistency of the data before feeding it into machine learning models. The following preprocessing steps were undertaken:

1. Data Cleaning: Missing values were handled using various imputation techniques. For instance, linear interpolation was used for continuous variables, while mode imputation was applied to categorical variables (Kotsiantis, Kanellopoulos, & Pintelas, 2006).
2. Feature Selection: Features were selected based on their correlation with air quality index (AQI). Highly correlated features were retained, while redundant and irrelevant features were removed to avoid multicollinearity (Guyon & Elisseeff, 2003).
3. Normalization: To ensure uniformity and improve the model's performance, the data was normalized using min-max scaling, transforming the feature values to a range between 0 and 1 (Han, Kamber, & Pei, 2012).

## 3.5 Machine Learning Models

Several machine learning algorithms were considered for predicting air quality. These models were selected based on their proven efficacy in regression tasks:

1. Linear Regression: A baseline model to understand the relationship between independent variables and Health Impact score. Despite its simplicity, linear regression provides a benchmark for comparing more complex models (Seber & Lee, 2012).
2. Random Forest Regression: An ensemble learning method that constructs multiple decision trees and outputs the mean prediction of the individual trees. Random Forests are robust to overfitting and can handle large datasets efficiently (Breiman, 2001).
3. Gradient Boosting Machines (GBM): GBM builds models sequentially, with each model attempting to correct the errors of its predecessor. This method is known for its high accuracy and efficiency in handling complex datasets (Friedman, 2001).
4. Support Vector Regression (SVR): SVR aims to find a hyperplane that best fits the data by maximizing the margin within a specified tolerance level. SVR is effective in high-dimensional spaces (Smola & Schölkopf, 2004).

## 3.6 Model Training and Evaluation

The dataset was split into training (70%) and testing (30%) sets to evaluate the models' performance. Each model was trained on the training set using 5-fold cross-validation to minimize overfitting and ensure generalization (Kohavi, 1995). The following metrics were used to evaluate the models:

1. Mean Absolute Error (MAE): Measures the average magnitude of errors in predictions without considering their direction (Willmott & Matsuura, 2005).
2. Root Mean Squared Error (RMSE): Provides a quadratic mean of the errors, which penalizes larger errors more significantly (Chai & Draxler, 2014).
3. R-squared (R²): Indicates the proportion of the variance in the dependent variable that is predictable from the independent variables (Nagelkerke, 1991).

## 3.7 Tools and Technologies

The implementation was carried out using Python programming language due to its extensive libraries and frameworks for machine learning. Key libraries included:

1. Pandas: For data manipulation and analysis.
2. NumPy: For numerical operations.
3. Scikit-learn: For implementing and evaluating machine learning models.
4. Matplotlib and Seaborn: For data visualization.

## 3.8 Conclusion

This chapter provided a comprehensive overview of the methodology used in predicting air quality using machine learning models. The steps involved in data collection, preprocessing, model selection, training, and evaluation were discussed in detail. The next chapter will present the results and analysis of the models' performance.

## References

Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.

Chai, T., & Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature. Geoscientific Model Development, 7(3), 1247-1250.

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

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. Journal of machine learning research, 3(Mar), 1157-1182.

Han, J., Kamber, M., & Pei, J. (2012). Data Mining: Concepts and Techniques. Elsevier.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In International Joint Conference on Artificial Intelligence (Vol. 14, No. 2, pp. 1137-1145).

Kotsiantis, S., Kanellopoulos, D., & Pintelas, P. (2006). Handling imbalanced datasets: A review. GESTS International Transactions on Computer Science and Engineering, 30(1), 25-36.

Nagelkerke, N. J. D. (1991). A note on a general definition of the coefficient of determination. Biometrika, 78(3), 691-692.

Seber, G. A. F., & Lee, A. J. (2012). Linear Regression Analysis. John Wiley & Sons.

Smola, A. J., & Schölkopf, B. (2004). A tutorial on support vector regression. Statistics and computing, 14(3), 199-222.

Willmott, C. J., & Matsuura, K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Climate Research, 30(1), 79-82.

Mishra, A. K., Goyal, P., & Shukla, A. K. (2023). Machine learning for air quality prediction: A comprehensive review. Atmospheric Pollution Research, 14(2), 101434. <https://www.sciencedirect.com/science/article/abs/pii/S1352231022004125>

US Environmental Protection Agency. (2023, June 22). Air Quality Index (AQI) - A Guide for Public Reporting. <https://www.airnow.gov/>

Wang, Y., Wang, Z., Li, S., Liu, Z., & Zhang, J. (2022). Deep learning for air quality prediction using satellite remote sensing data. Atmospheric Environment (X), 100080. <https://www.sciencedirect.com/science/article/abs/pii/S1352231022004125>