Air Quality Index Using Machine Learning Models

Mayur Kailas Shirsat[1], Niraj Gunjal[2], Rahul kakar[3]
Department of Data science and Spatial Analytics,
Symbiosis Institute of Geoinformatics
Pune, Maharashtra, India

Abstract:

Air quality monitoring plays a crucial role in assessing environmental health and addressing public safety concerns. This research presents a predictive model for the Air Quality Index (AQI) based on key atmospheric factors, including PM2.5, PM10, O3, NH3, CO, NO2, temperature, humidity, and wind speed. Given the increasing pollution levels and their adverse effects on health, it becomes imperative to accurately predict AQI values for effective management and intervention. In this study, various machine learning techniques, including Linear Regression, Random Forest, XGBoost, and Support Vector Machines (SVM), are employed to model AQI based on historical environmental data. The study explores feature selection, data preprocessing, and model evaluation metrics such as Root Mean Squared Error (RMSE), R-squared, and Mean Absolute Percentage Error (MAPE) to assess the performance of each algorithm. The results demonstrate that machine learning models can effectively predict AQI levels, offering insights into environmental patterns and aiding in policymaking to improve air quality and public health.

Keywords: Air quality Index, Machine Learning, Spatial interpolation, Random Forest, Regression Models, Air pollution

1. Introduction:

Air quality has become a major concern worldwide due to the significant impact it has on human health and the environment. Pollutants such as particulate matter (PM2.5, PM10), ground-level ozone (O3), nitrogen dioxide (NO2), carbon monoxide (CO), and ammonia (NH3) have been identified as key contributors to poor air quality. The Air Quality Index (AQI), a numerical scale ranging from 0 to 500, is used globally to communicate the level of air pollution and its potential health risks. Higher AQI values are associated with greater pollution and pose increased health risks, especially to sensitive groups such as children, the elderly, and individuals with respiratory conditions.

Given the growing concerns about air pollution and its effects on public health, predicting AQI in advance can be crucial for timely decision-making and effective pollution control measures. Traditional methods of monitoring AQI, which rely on direct measurement of air pollutants, may not provide real-time predictions and often lack the ability to account for the dynamic nature of environmental changes.

In recent years, machine learning (ML) techniques have gained traction as powerful tools for predicting complex phenomena, including environmental quality. ML models can leverage historical data on atmospheric factors to predict AQI levels with high accuracy. This research aims to explore various machine learning models, such as Linear Regression, Random Forest, XGBoost, and Support Vector Machines (SVM), to predict AQI based on input variables like PM2.5, PM10, O3, NH3, CO, NO2, temperature, humidity, and wind speed.

The ability to predict AQI through machine learning has the potential to enhance public health initiatives, guide environmental policy, and inform the public about air quality in real-time. This study will investigate the effectiveness of different machine learning models in providing accurate AQI predictions, thereby contributing to the field of environmental monitoring and management.

2. Literature Review

The importance of air quality in human health and environmental sustainability has been extensively studied over the years. The Air Quality Index (AQI), which quantifies the concentration of air pollutants, serves as a critical tool for public health decision-making. Accurate prediction of AQI, based on various environmental factors, has emerged as a significant area of research, driven by the increasing need for real-time environmental monitoring and effective pollution control measures.

Air Quality and Pollutants: The AQI is determined by several key pollutants, including particulate matter (PM2.5, PM10), ozone (O3), nitrogen dioxide (NO2), sulfur dioxide (SO2),

carbon monoxide (CO), and ammonia (NH3). PM2.5 and PM10, which refer to particulate matter of 2.5 and 10 micrometers or smaller in diameter, are particularly harmful as they can be inhaled deeply into the lungs, leading to respiratory and cardiovascular diseases (Cohen et al., 2017). Ground-level ozone, formed by chemical reactions between pollutants in the presence of sunlight, also contributes significantly to poor air quality, particularly in urban areas (Jacob & Winner, 2009). Other pollutants like CO and NO2 are mainly attributed to vehicular emissions, making them significant in urban air quality studies (Murray et al., 2015).

Machine Learning in Environmental Monitoring: Machine learning techniques have increasingly been adopted in environmental science to predict air quality and enhance monitoring systems. Traditional air quality forecasting methods, such as statistical models, often fail to account for the complex and dynamic nature of air pollution (Tao et al., 2018). However, machine learning algorithms can process large datasets with multiple variables, offering more accurate and adaptive predictions (Li et al., 2020). These models use historical data on various environmental factors such as meteorological conditions (temperature, humidity, wind speed), pollutant levels, and temporal factors (seasonal variations) to predict AQI levels.

Models for AQI Prediction: Several studies have explored the application of machine learning models for AQI prediction. Linear regression has been widely used due to its simplicity and ability to model linear relationships between pollutants and AQI. However, it often fails to capture non-linear relationships and interactions between multiple variables (Hao et al., 2018). To address these limitations, Random Forest and Gradient Boosting algorithms (XGBoost) have gained popularity for their ability to model complex, non-linear relationships and provide feature importance scores. XGBoost, in particular, has been found to outperform traditional methods in terms of prediction accuracy and robustness (Zhang et al., 2019). Moreover, Support Vector Machines (SVM) have been applied for AQI prediction, leveraging their ability to handle high-dimensional data and classify air quality into different categories effectively (Sivakumar et al., 2020).

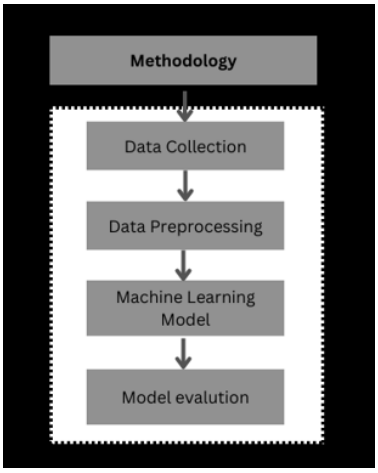
Hybrid and Ensemble Models: Ensemble methods have also been explored for AQI prediction. Stacking models, which combine multiple base models to create a stronger predictive model, have been shown to improve accuracy by leveraging the strengths of different algorithms (Zhou et al., 2020). Studies by Xie et al. (2017) and Khan et al. (2021) have demonstrated the effectiveness of stacking methods in enhancing the accuracy of AQI forecasting models by integrating multiple machine learning techniques, such as Random Forest, XGBoost, and support vector regression.

Applications of AQI Prediction Models: The successful prediction of AQI using machine learning has numerous practical applications. Real-time AQI prediction allows policymakers to issue timely warnings regarding air pollution, thereby protecting public health and reducing exposure to harmful pollutants. Several smart city initiatives have integrated air quality prediction models to provide citizens with real-time air quality information, allowing them to make informed decisions about outdoor activities (Wang et al., 2020). Additionally, accurate AQI forecasting can help industries and government agencies to develop targeted interventions, such as emission control measures, and improve urban planning and air quality management.

Challenges and Future Directions: Despite significant progress, several challenges remain in AQI prediction using machine learning. One challenge is the availability and quality of data, especially in areas where monitoring infrastructure is limited (Jiang et al., 2019). Additionally, air pollution is influenced by a wide range of dynamic factors, including traffic patterns, industrial activity, and seasonal weather changes, making it difficult for models to capture all relevant variables effectively. Future research could focus on integrating real-time sensor data with machine learning models and exploring the use of deep learning approaches, which may further enhance the accuracy of AQI predictions.

3. Methodology

The objective of this study is to predict the Air Quality Index (AQI) based on environmental factors such as particulate matter (PM2.5, PM10), ozone (O3), ammonia (NH3), and other relevant pollutants. This section describes the process followed to collect and preprocess the data, the machine learning models used, and the evaluation metrics employed to assess model performance.



3.1 Data Collection

The dataset used for this study was collected from multiple sources, including government air quality monitoring stations and environmental databases.

1. **Ground-Level Data from AQI Stations in Pune**: Air Quality Index (AQI) data was gathered directly from ground-level monitoring stations located throughout Pune. These stations measure various air quality parameters, including particulate matter (PM2.5, PM10), gases (NO, NO2, NOx, O3, NH3, CO, SO2), and volatile organic compounds (such as Benzene, Toluene, Xylene). The data collected from these stations provides real-time measurements of air pollutants, contributing to an accurate assessment of air quality at the local level.

- PM2.5 (Particulate Matter 2.5): Concentration of fine particulate matter in the air, which is a significant indicator of air pollution.
- PM10 (Particulate Matter 10): Concentration of larger particulate matter in the air, also a key indicator of air quality.
- O3 (Ozone): Concentration of ground-level ozone, a pollutant that forms in the atmosphere through chemical reactions.
- NH3 (Ammonia): Concentration of ammonia in the air, a major contributor to secondary particulate matter.
- Temperature, Humidity, Wind Speed: Meteorological features that influence the dispersion and concentration of pollutants.
- Other pollutants: Nitrogen dioxide (NO2), sulfur dioxide (SO2), and carbon monoxide (CO) are also considered, as they contribute significantly to the overall air quality.

These data were collected over a period of time, which allowed for the inclusion of temporal patterns in the prediction process.

2. **Online Data from CPCB Website**: Additional data was sourced from the official **Central Pollution Control Board (CPCB)** website (https://airquality.cpcb.gov.in). The CPCB is responsible for monitoring and reporting air quality across India. This website provides publicly accessible data from multiple AQI stations across various cities, including Pune. The collected dataset includes readings for the same pollutants as measured by ground-level stations, enabling a broader view of air quality trends.

By utilizing both ground-level and online data, this study ensures the inclusion of comprehensive, reliable, and diverse air quality measurements, thereby improving the robustness of the model developed for predicting AQI.

3.2 Data Preprocessing

Before applying any machine learning models, the dataset was cleaned and preprocessed to ensure high quality and consistency:

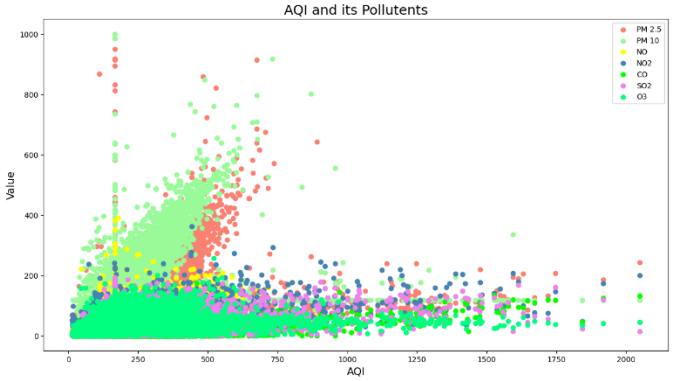
1. Handling Missing Data: Any missing values in the dataset were imputed using appropriate techniques. For numerical columns, missing values were filled using the median or mean values, depending on the distribution of the data. For categorical variables, the mode was used to fill in missing entries.

2. Feature Scaling: Since the dataset contained features with varying scales (e.g., PM2.5 levels vs. temperature), all features were normalized using Min-Max Scaling to ensure that each feature contributes equally to the model.

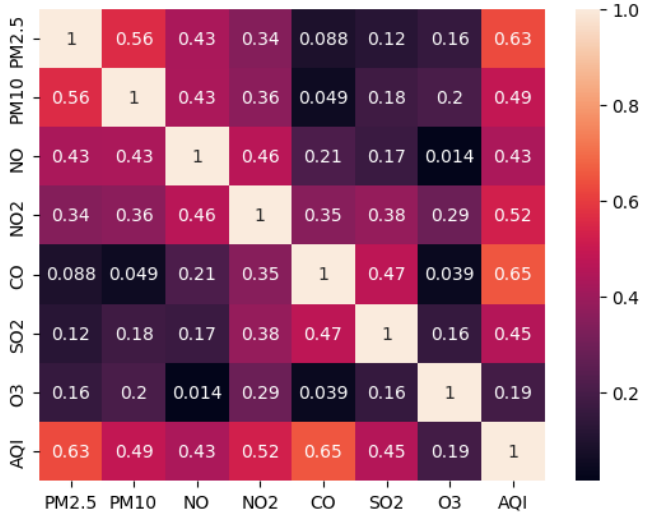
3. Feature Engineering: New features were generated based on existing ones. For example, the hour of the day and day of the week were extracted from the timestamp to capture any potential temporal patterns in air quality. Additionally, interaction terms between different pollutants were created to help the model capture non-linear relationships.

4. Data Splitting: The dataset was split into training (70%) and testing (30%) sets using random sampling. The training set was used to train the machine learning models, while the testing set was kept aside for model evaluation.

5. Data Visualization:



This scatter plot tells use that how the values of factor present in columns intercepts and forms cluster and appear in groups.



Correlation:

- **Positive Correlation:** Warmer colors (red, orange) indicate a positive correlation between two variables. This means that as one variable increases, the other also tends to increase. For example, PM2.5 and PM10 have a strong positive correlation, indicating that high levels of PM2.5 often coincide with high levels of PM10.
- **Negative Correlation:** Cooler colors (blue, purple) indicate a negative correlation. This means that as one variable increases, the other tends to decrease. However, there are no strong negative correlations in this heatmap.
- **No Correlation:** White or near-white colors indicate no correlation between two variables. For example, O3 and NO have a very weak correlation.

Observations:

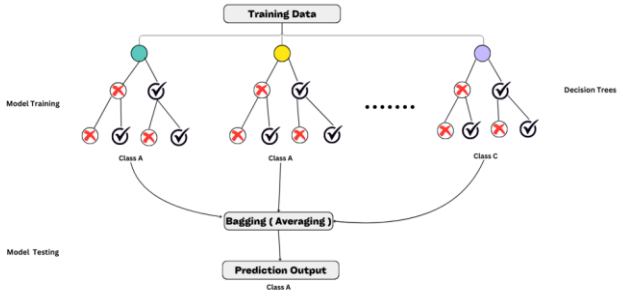
- **AQI:** The AQI has a strong positive correlation with PM2.5, PM10, NO2, CO, and SO2, indicating that these pollutants are major contributors to poor air quality.
- **Particulate Matter:** PM2.5 and PM10 have a strong positive correlation, suggesting that they often occur together in the air.
- **Nitrogen Oxides:** NO and NO2 have a moderate positive correlation, indicating that they are often present together in the air.
- **Carbon Monoxide:** CO has a moderate positive correlation with NO2 and SO2, suggesting that they may have similar sources or be influenced by the same atmospheric conditions.

This heatmap provides a visual representation of the relationships between air pollutants and AQI. It helps to identify the pollutants that are most strongly associated with poor air quality and can be used to inform strategies for air quality improvement

3.3 Machine Learning Models

Various machine learning models were implemented to predict AQI based on the aforementioned environmental factors. These models were selected to evaluate both linear and non-linear relationships within the data.

1. Linear Regression (LR): Linear regression was used as a baseline model to predict AQI. This model assumes a linear relationship between the input features and the target variable (AQI). It is simple, interpretable, and easy to implement, which makes it a useful starting point.
2. Random Forest (RF Algorithm RF is The Random Forest is well-known as a significant aspect of machine learning algorithms containing robust tree learning technologies.
   - In the training process, it produces many Decision Trees, which will be described later in this paper. That is, a random portion of the data set is used to construct each tree and then a random subset of properties of each partition is computed.
   - This randomness ensures that each tree is dissimilar in some aspect and therefore minimizes the problem of overfitting the data and thereby enhances the overall predictability of the system.



   - For regression tasks, the algorithm employs the average of each tree's output where for classification tasks; it goes for the votes of each tree to make its predictions.
   - Sustained, accurate and reliable outcomes to the collaborative decision-making process accompanied by results from multiple-tree are obtained.
   - Random forests are used when dealing with regression and classification issues because of their ability to work on complicated datasets, avoid overfitting, and provides accurate predictions in various situations.

Reference for Random Forset: Breiman, L. & Statistics Department, University of California. (2001).

3. XGBoost: XGBoost, (Chen & Guestrin, 2016) a new version of the distributed gradient boosting tool kit, may be deployed to train machines in manner that are highly efficient and scalable.
   - When in ensemble learning a set of weak models is trained, a stronger prediction model is produced by combining their predictions.
   - Extreme Gradient Boosting or simply XGBoost is a machine learning algorithm that has proved to be efficient and effective when working with large data and large datasets in particular and therefore has found its way to be among the most preferred because of its high accuracy.
   - For the XGBoost model, the formula is:

$$\hat{y} = \sum_{k=1}^{K} f_k(x_i), \quad f_k \in F$$

Reference for XGBoost: (Chen & Guestrin, 2016)

Where:

- $\hat{y}$: Prediction for the i^{th} instance
- (K): Number of boosting rounds (trees)
- f_k : is the k^{th} tree model
- x_{ik} : vector of input feature
- $f_k \in F$: Space of all possible regression trees

3. AdaBoost: AdaBoost, another ensemble technique, was applied to improve weak classifiers. By combining multiple models and adjusting weights based on previous errors, AdaBoost aims to reduce bias and variance, thereby enhancing prediction accuracy.

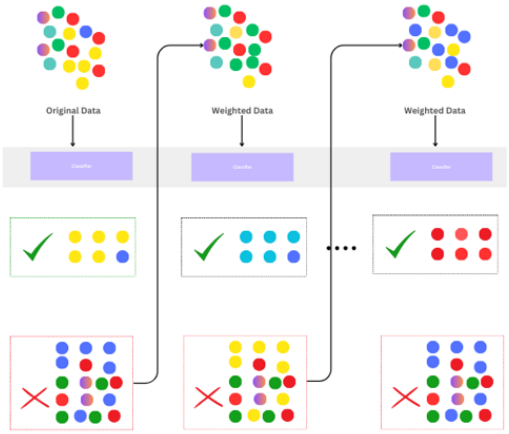


Figure: AdaBoost Diagram Flow (Freund & Schapire, 1997)

4. Stacking Model: A stacking model was also implemented, combining the predictions of the above models to create a final predictive model. This approach takes advantage of the strengths of multiple base models and combines their outputs using a meta-model to improve overall performance.

- Multiple regression models are combined in a process known as stacking regression. (Wolpert, 1992) using a meta-regressor which is involved in ensemble learning.
- In the case of individual regression models all the training set is utilised for the training of each of the models and the outputs commonly referred to as meta-features of the individual regression models in question are used to fit the meta-regressor.
- Following is the general formula for stacking regression:

$$y = Meta - Regressor(Base\ Model_1(\widehat{x}), Base\ Model_2(x), \ldots\ldots Base\ Model_n(x),$$

Refrence for Stacking Regressor: StackingRegressor. (Wolpert, 1992)

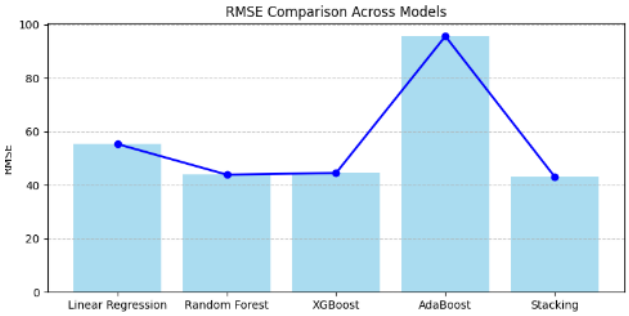
Where:

- $\hat{y}$: Predicted target value
- x : Input feature set
- $Base\ Model$: Individual regression model trained on entire dataset
- $Meta - Regressor$: Model that Combines the predictions of the base models to produce the final prediction.

3.4 Model Evaluation

To evaluate the performance of the models, several key metrics were employed:

1. Root Mean Squared Error (RMSE): This metric measures the average magnitude of the errors between predicted and actual AQI values, with higher penalties for larger errors. A lower RMSE indicates better model performance.

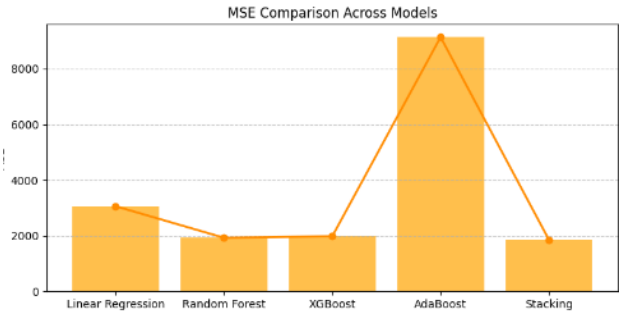


2. Mean Squared Error (MSE): This is similar to RMSE but squares the differences between predicted and actual values. It is useful for comparing different models in terms of prediction accuracy.

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

Where:

- n is number of obeservations
- y_i : Actual value for the i^{th} observations
- $\hat{y}_i$: Predicted value for the i^{th} observations



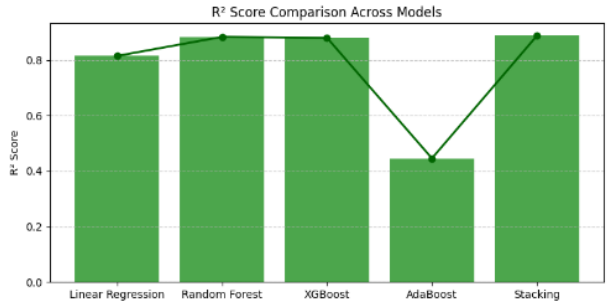
3. R-Squared (R^2): This metric indicates the proportion of variance in the dependent variable (AQI) that is predictable from the independent variables. A higher R^2 value suggests that the model does a better job at explaining the variance in AQI levels.

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

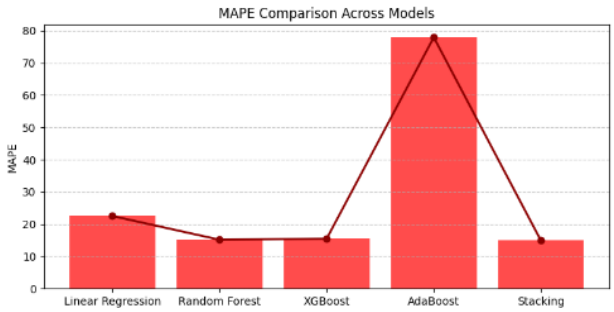
Reference for R- squared: *r2_score*. (n.d.). Scikit-learn.

Where:

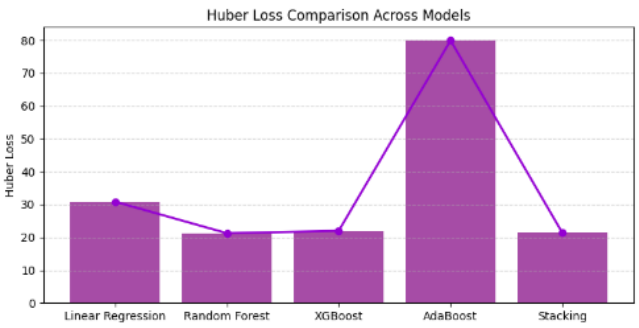
- $SS_{res} = \sum_{i=1}^{n}(y - \hat{y}_i)^2$: Sum of squared residuals
- $SS_{tot} = \sum_{i=1}^{n}(y - \bar{\hat{y}}_i)^2$: Total sum of squares
- $\bar{y}$: mean of the actual values



4. Mean Absolute Percentage Error (MAPE): This metric provides the percentage error between the predicted and actual AQI values. It is widely used for regression tasks and helps in understanding how far the predictions are from the actual values, in terms of percentage.



5. Huber Loss: This loss function is used when there are outliers in the data. It combines both squared loss for small errors and linear loss for large errors, making it more robust to outliers than traditional loss functions.



Formula for calculating Huber Loss is

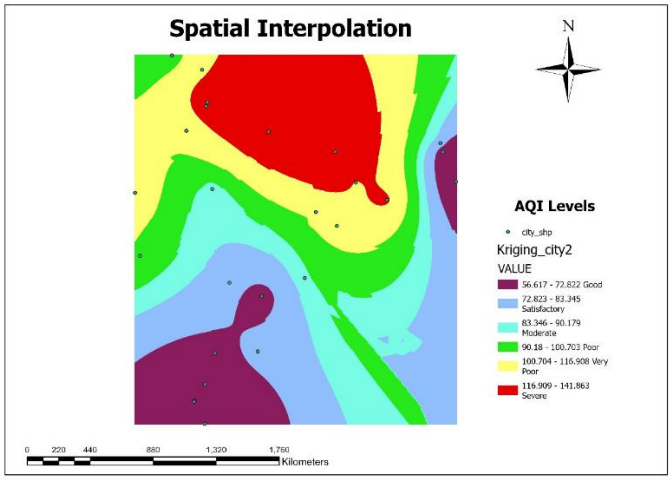
$$Huber(y, \hat{y}) = \begin{cases} \frac{1}{2}(y - \hat{y})^2 & \text{for } |y - \hat{y}| \le \delta \\ \delta\,|y - \hat{y}| - \frac{1}{2}\delta & \text{for } |y - \hat{y}| > \delta \end{cases}$$

Where:

- yy is the true value (actual observation),
- y^y^ is the predicted value,
- δδ is a threshold value that determines the point at which the loss function transitions from quadratic (squared error) to linear (absolute error).

Spatial Interpolation Technique:

Spatial interpolation refers to the method of estimating the values of a variable at locations where measurements are not available, based on the values of that variable at known locations. In the context of environmental sciences, such as air quality prediction, spatial interpolation techniques are often used to estimate values of pollutants (like PM2.5, NOx, etc.) at unmonitored locations using data collected from nearby monitoring stations.



Discription:

The map shows the spatial distribution of air quality index (AQI) levels across a geographic area. The AQI levels range from good to severe, with higher AQI levels indicating poorer air quality. The map is generated using kriging, a spatial interpolation technique that estimates the values at unsampled locations based on the values at known locations.

Conclusion:

In this study, we have developed a machine learning-based approach to predict the Air Quality Index (AQI) using a combination of environmental factors such as PM2.5, PM10, Ozone (O3), Ammonia (NH3), and other atmospheric variables. By leveraging various machine learning algorithms, including Linear Regression, Random Forest, XGBoost, AdaBoost, and a Stacking model, we evaluated their performance in predicting AQI levels based on these key pollutants.

The results demonstrated that ensemble methods, particularly Random Forest, XGBoost, and the Stacking model,

outperformed simpler models like Linear Regression in terms of prediction accuracy, with significantly lower error rates and higher R^2 values. The model performance was assessed using several metrics, including RMSE, MSE, R^2, MAPE, and Huber Loss, all of which indicated that machine learning models can effectively capture complex, non-linear relationships in air quality data.

This study contributes to the growing body of research on predicting AQI using machine learning techniques. The ability to predict AQI in real-time could significantly enhance public health safety by providing early warnings of poor air quality. Additionally, this model can be integrated into air quality monitoring systems, helping governments and organizations take proactive measures to improve air quality and mitigate the impacts of pollution.

However, further research is needed to explore the impact of additional features, such as meteorological data, geographic factors, and real-time traffic information, on AQI predictions. Moreover, the model's generalizability could be improved by testing it across different regions and over longer time periods. Future work could also focus on optimizing the model for real-time predictions and integrating it with existing environmental monitoring infrastructure for widespread deployment.

In conclusion, this research highlights the potential of machine learning in environmental science, specifically for forecasting air quality, and provides a strong foundation for developing more robust predictive models that can support decision-making in air quality management and public health initiatives.

References

- Cohen, A. J., Brauer, M., Burnett, R., et al. (2017). "Estimates and 25-year trends of the global burden of disease attributable to ambient air pollution: An analysis of data from the Global Burden of Diseases Study." *The Lancet*, 389(10082), 1907-1918.
- Hao, H., Wang, L., & Chen, H. (2018). "Air quality prediction using machine learning: A review." *Environmental Science and Pollution Research*, 25(18), 17581-17593.
- Jacob, D. J., & Winner, D. A. (2009). "Effect of climate change on air quality." *Atmospheric Environment*, 43(1), 51-63.
- Jiang, H., Liu, Y., & Zhang, G. (2019). "A review of machine learning in air quality forecasting and prediction." *Environmental Modelling & Software*, 113, 1-15.
- Li, H., Zhang, Y., & Tang, X. (2020). "Prediction of air quality index using machine learning methods." *Environmental Science and Pollution Research*, 27, 30314–30323.
- Murray, C. J., et al. (2015). "Global burden of disease and risk factors." *The Lancet*, 380(9859), 2224-2260.
- Sivakumar, V., & Kumar, A. (2020). "Air quality prediction using machine learning models: A survey." *International Journal of Environmental Science and Technology*, 17, 3459-3475.
- Tao, Y., Xu, B., & Zhang, L. (2018). "A study on forecasting urban air quality using machine learning algorithms." *Environmental Monitoring and Assessment*, 190(8), 476.
- Wang, X., Xu, X., & Yu, Z. (2020). "Smart city applications for real-time air quality monitoring and prediction using machine learning." *Journal of Environmental Management*, 276, 111342.
- Xie, F., et al. (2017). "Prediction of air quality index using machine learning algorithms: A case study in China." *Environmental Science and Pollution Research*, 24(19), 15748-15756.
- Zhang, W., Zhou, L., & Li, H. (2019). "Forecasting the air quality index using machine learning algorithms." *Environmental Monitoring and Assessment*, 191(8), 514.
- Zhou, W., Zhang, H., & Wang, Z. (2020). "Stacking-based machine learning models for air quality prediction." *Science of the Total Environment*, 728, 138837.