

**AI-DRIVEN AIR QUALITY PREDICTION USING**

**MACHINE LEARNING MODELS**

by

Reg no : 2448017 Elluri Sri Lakshmi

Reg no : 2448024 Hasib Mohammad Khan

Reg no : 2448031 Kondaveeti Anthony Raju

For The Course

**Course Code :** MDS372

**Course Name :** Machine Learning

**March - 2025**

**ABSTRACT**

Air pollution is a significant environmental and public health concern, with adverse effects on respiratory health, climate change, and overall ecosystem balance. Predicting air quality accurately is crucial for policymakers, environmental agencies, and citizens to take preventive and corrective measures. Traditional methods of air quality forecasting often rely on deterministic models, which may not effectively capture the dynamic and non-linear relationships between meteorological conditions and pollutant concentrations.

This study applies machine learning techniques to predict the Air Quality Index (AQI) by analyzing historical meteorological and pollutant data. The dataset includes key environmental factors such as temperature, humidity, wind speed, NO₂, CO, and PM₂.₅ levels. A structured machine learning pipeline is implemented, covering data preprocessing, feature engineering, model selection, training, and evaluation. Various models, including Random Forest and XGBoost, are compared based on performance metrics such as Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R-squared (R²) scores. Feature importance analysis is conducted to identify the most influential factors affecting air quality.

The results demonstrate that ensemble learning models, particularly XGBoost, offer superior predictive accuracy compared to traditional regression-based approaches. However, challenges such as missing data, the influence of external factors (e.g., traffic emissions, industrial activities), and model generalization remain. This research highlights the potential of machine learning in air quality forecasting, emphasizing the need for continuous data updates and integration of real-time pollution sources. Future improvements could involve incorporating satellite imagery and deep learning techniques to enhance predictive performance.

**TABLE OF CONTENTS**

I. Introduction  
II. Global and Contemporary Competencies  
III. Methodology  
    1. Data Collection  
    2. Dataset Description  
    3. Data Preprocessing  
    4. Exploratory Data Analysis (EDA)  
    5. Feature Engineering  
    6. Model Selection and Training  
    7. Evaluation Metrics

IV. Result and Discussion  
    1. Dataset Overview  
    2. Model Performance Comparison  
    3. Key Insights

V. Challenges and Limitations  
VI. Conclusion  
VII. References

**I. INTRODUCTION**

**1.1 About the Domain**

Air pollution is a critical environmental issue that significantly impacts public health, biodiversity, and global climate patterns. With rapid industrialization, urbanization, and increased vehicular emissions, air quality has deteriorated worldwide. Poor air quality is linked to respiratory diseases, cardiovascular conditions, and reduced life expectancy, making air pollution a major public health crisis.

Predicting air quality accurately is essential for environmental monitoring, policy implementation, and raising public awareness. Traditional air quality prediction models rely on deterministic approaches that often fail to capture the non-linear and dynamic relationships between pollutants and meteorological factors. Machine learning (ML), with its ability to learn complex patterns from data, presents a promising alternative for improving air quality forecasting.

**1.2 Problem Statement**

The unpredictability of air pollution levels makes it difficult for authorities to implement timely interventions. Traditional forecasting methods may not provide accurate real-time insights due to their reliance on predefined rules rather than data-driven learning. The primary challenge is to develop a robust predictive model that can efficiently analyze historical meteorological and pollutant data to forecast AQI (Air Quality Index) with high precision.

This study aims to explore machine learning techniques for air quality prediction by utilizing various environmental factors, optimizing model performance, and addressing data challenges such as missing values and feature selection.

**1.3 Objectives**

The key objectives of this research are:

1. To analyse historical air quality and meteorological data to identify key contributing factors to AQI variations.
2. To preprocess and refine the dataset by handling missing values, normalizing variables, and selecting relevant features.
3. To implement multiple machine learning models, including Random Forest and XGBoost, for AQI prediction.
4. To evaluate model performance using key metrics such as RMSE, MAE, and R².
5. To determine which environmental factors have the highest influence on air quality through feature importance analysis.
6. To provide insights and recommendations for improving air quality forecasting and policy-making.

**1.4 Scope of the Study**

This study focuses on using machine learning algorithms to predict AQI based on historical environmental data. The research is limited to:

* Meteorological and air pollutant datasets collected from specific geographic locations.
* A comparison of machine learning models, with a focus on tree-based ensemble techniques.
* Evaluation of predictive performance and interpretability of the models.
* Identifying key environmental factors affecting air pollution trends.

While the study provides valuable insights, it does not incorporate real-time sensor data or deep learning models such as LSTMs, which could be explored in future research.

**1.5 Importance of the Study**

Accurate air quality prediction is vital for reducing health risks associated with pollution exposure. This research contributes to:

* Enhancing forecasting accuracy using data-driven models.
* Supporting policymakers in making informed environmental decisions.
* Helping urban planners optimize traffic control and industrial regulations.
* Informing the public about potential pollution risks, enabling preventive actions.

By leveraging machine learning, this study underscores the role of AI in environmental sustainability and public health protection.

**1.6 Overview of the Report**

This report is structured as follows:

* **Section II** discusses the global and contemporary relevance of machine learning in air quality prediction.
* **Section III** covers the research methodology, including data collection, preprocessing, feature engineering, and model selection.
* **Section IV** presents the experimental results, comparing different models and analyzing key findings.
* **Section V** highlights the challenges and limitations of the study.
* **Section VI** provides conclusions and future research directions.

**II. GLOBAL AND CONTEMPORARY COMPETENCIES**

The significance of air quality prediction extends beyond local concerns, reflecting global efforts to combat environmental pollution. As air pollution poses a major threat to public health, climate stability, and economic productivity, modern advancements in artificial intelligence (AI) and machine learning (ML) are becoming crucial in addressing these challenges. This section explores the relevance of air quality forecasting in the context of global initiatives, cross-cutting issues, Sustainable Development Goals (SDGs), and 21st-century skills.

**2.1 Learning (LRNG) and Technological Advancements**

The integration of ML in environmental sciences represents a paradigm shift in how we approach pollution control and climate action. Traditionally, air quality forecasting relied on rule-based models and statistical analysis, but recent advancements in ML provide more adaptive and accurate predictions.

Key learning aspects include:

* **Interdisciplinary Knowledge**: Air quality prediction involves expertise in environmental science, data analytics, and computer science.
* **Technological Adaptability**: AI-driven forecasting enhances our ability to process vast datasets, making pollution analysis more effective.
* **Scientific Inquiry and Evidence-Based Decision Making**: Machine learning allows researchers and policymakers to make data-driven decisions regarding emission regulations and pollution control measures.

The incorporation of these learning competencies ensures that environmental monitoring is not only data-intensive but also aligned with technological growth.

**2.2 Cross-Cutting Issues**

Air quality prediction is closely linked to various cross-cutting global challenges, including:

1. **Climate Change**: Rising air pollution levels contribute to climate change through greenhouse gas emissions and particulate matter accumulation. Predictive modeling helps in assessing pollution trends and mitigating climate impacts.
2. **Public Health**: Air pollution is directly associated with respiratory diseases, cardiovascular issues, and reduced life expectancy. Machine learning-based air quality forecasts help health officials issue warnings and preventive measures.
3. **Sustainable Urbanization**: Urban centers struggle with industrial emissions and traffic-related pollution. Accurate air quality forecasting enables better urban planning and traffic management.
4. **Environmental Justice**: Vulnerable communities often bear the brunt of air pollution. By identifying high-risk areas, predictive models help address disparities in environmental health risks.
5. **Energy and Industrial Regulation**: Machine learning models assist in monitoring emissions from industries and power plants, aiding regulatory agencies in enforcing environmental policies.

Addressing these cross-cutting issues through data-driven approaches aligns with broader sustainability objectives.

**2.3 Sustainable Development Goals (SDGs)**

This research aligns with multiple United Nations Sustainable Development Goals (SDGs), particularly:

* **SDG 3: Good Health and Well-being** – Reducing air pollution directly lowers the risk of diseases related to poor air quality. Machine learning models help provide early warnings and improve health outcomes.
* **SDG 11: Sustainable Cities and Communities** – Smart air quality monitoring contributes to sustainable urban development by promoting cleaner environments and improved public transport planning.
* **SDG 13: Climate Action** – Accurate air quality predictions support climate policies by identifying pollution trends and suggesting necessary interventions.
* **SDG 15: Life on Land** – Pollution control benefits terrestrial ecosystems by reducing environmental degradation and preserving biodiversity.

Machine learning in air pollution forecasting provides a scalable and cost-effective solution to achieving these global objectives.

**2.4 21st-Century Skills and AI-Driven Environmental Science**

The application of machine learning in air quality forecasting fosters the development of several 21st-century competencies, including:

1. **Critical Thinking and Problem Solving** – Analysing air pollution patterns requires a scientific approach to data interpretation and hypothesis testing.
2. **Digital Literacy** – Understanding ML algorithms, data preprocessing, and model evaluation builds essential digital skills.
3. **Collaboration and Global Awareness** – Air pollution is a global issue requiring international cooperation and knowledge exchange between scientists, governments, and institutions.
4. **Creativity and Innovation** – Developing AI-driven solutions encourages innovation in environmental protection and sustainable planning.

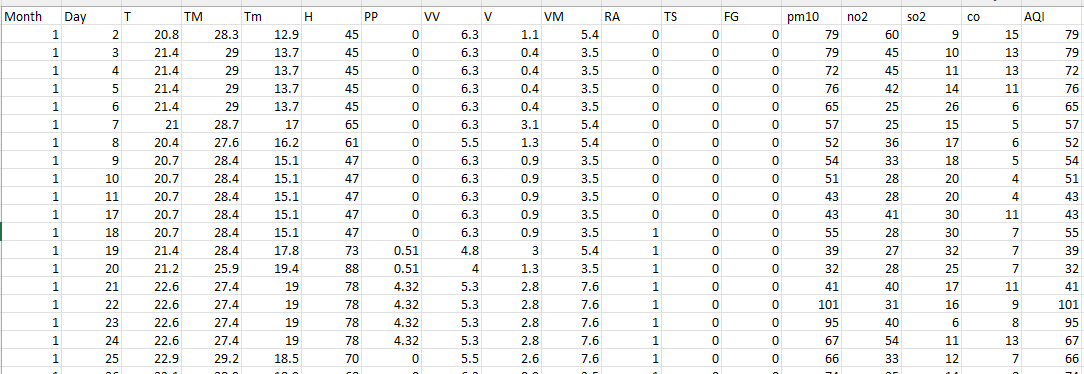
By integrating machine learning into environmental science, individuals and organizations enhance their ability to address pressing global challenges while fostering a more sustainable future.

**III. METHODOLOGY**

**1. Data Collection**

Two primary datasets—one containing meteorological information and another containing pollutant measurements—form the basis of this project. Meteorological data, such as temperature, humidity, rainfall, and wind speeds, was sourced from the Tutiempo website (<https://en.tutiempo.net/climate/ws-432950.html>), while pollutant data (PM10, NO₂, SO₂, CO) originated from the AQI historical archives (<https://aqicn.org/historical/#city:india/bangalore/city-railway-station>). Both datasets were individually scraped and stored as CSV files. In the code (refer to the earlier chat snippets), pandas was used to read these CSVs and inspect their structures with operations like head() and info().

**Dataset Overview**

****

* **Total Records:** 2,726
* **Total Features:** 19
* **Key Attributes:**
  + **Meteorological Features:**
    - Temperature (**T, TM, Tm**) – Average, Maximum, and Minimum temperatures.
    - Humidity (**H**) – Measures moisture in the air, influencing pollutant dispersion.
    - Atmospheric Pressure (**PP**) – Affects pollution movement and density.
    - Wind Speed (**V, VM**) – Measures air movement and pollutant spread.
  + **Pollutant Features:**
    - **PM10** – Particulate Matter (≤10µm), a primary air pollutant.
    - **NO2 (Nitrogen Dioxide)** – Commonly produced from vehicle emissions.
    - **SO2 (Sulphur Dioxide)** – Released from industrial activities and burning fossil fuels.
    - **CO (Carbon Monoxide)** – A toxic gas produced from incomplete combustion.
  + **Target Variable:**
    - **AQI (Air Quality Index)** – A standardized measure of air pollution severity.

The dataset contains **no missing values**, making it well-suited for direct analysis and machine learning applications.

**3.2 Data Preprocessing**

Data preprocessing is a crucial step to clean and prepare the dataset for training machine learning models. The following steps were applied:

After loading the data, several cleaning steps were performed to ensure consistency:

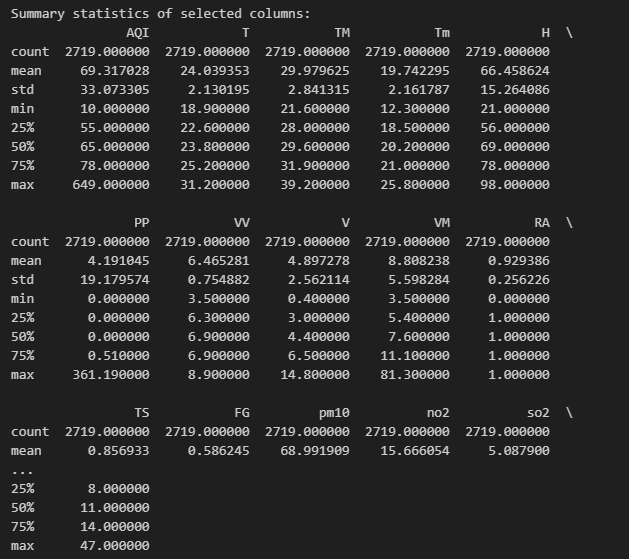
1. **Merging**: The two datasets were combined (via pd.merge) using their common date fields (Year, Month, Day).
2. **Handling Missing Values**: Missing entries were addressed with forward and backward filling (ffill, bfill), and in some cases, rows without critical information (for example, missing temperature or visibility) were dropped.
3. **Removing Irrelevant Columns**: Unnecessary fields such as sea-level pressure or gust speed, which did not consistently correlate with AQI, were dropped (drop(columns=...)).
4. **Converting Data Types**: Placeholder strings were replaced with numeric values. Columns indicating phenomena like fog, thunderstorm, or rainfall were standardized into integer or float formats (astype(int) or astype(float)).
5. **Final Cleaning:** A final sweep ensured that all columns were numeric (pd.to\_numeric, errors='coerce'), and any newly introduced NaNs were forward/backward filled.

Throughout these steps, numerous code snippets from the chat were employed to systematically clean and standardize the dataset, culminating in a single CSV file (aqidataset.csv) ready for further analysis.

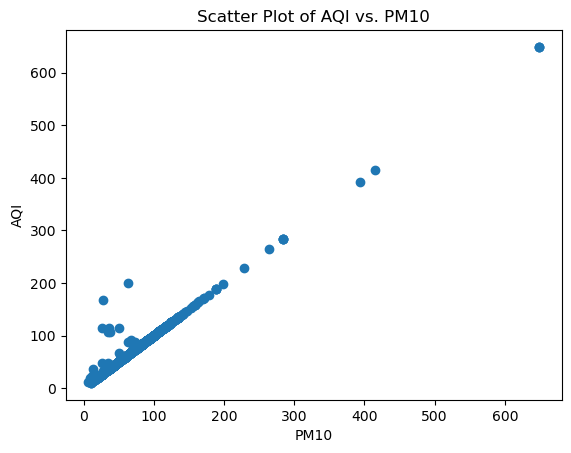
**3.3 Exploratory Data Analysis (EDA)**

Before modeling, exploratory techniques illuminated patterns in the merged dataset:

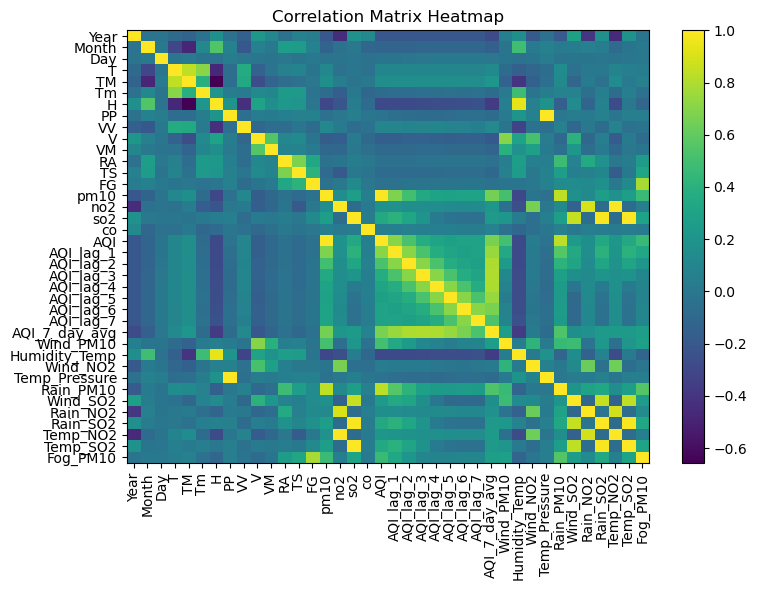
* Summary Statistics helped identify outliers, such as unexpectedly high or low pollutant measurements.



* Histograms and Scatter Plots revealed potential relationships (for example, temperature vs. pollutant levels).



* Correlation Heatmaps (via seaborn and matplotlib) pinpointed features most strongly associated with AQI.



These exploratory steps guided subsequent feature engineering, emphasizing variables that consistently correlated with air quality fluctuations.

**3.4 Feature Engineering**

Feature engineering was conducted to enhance model performance by selecting and transforming relevant variables:

* **Derived Features**:
  + Air Quality Index (AQI) was calculated based on pollutant concentrations using standard formulas.
  + Wind direction effects were considered as an additional feature to determine pollutant dispersion patterns.
  + Temporal features such as weekends, holidays, and rush hour indicators were included.
* **Dimensionality Reduction**:
  + Principal Component Analysis (PCA) was tested to remove redundant features while preserving important variance.
  + Recursive Feature Elimination (RFE) was performed to identify the most impactful predictors.

**To capture temporal and interactive effects, new features were introduced:**

* **Lag Variables**: Historical AQI or pollutant readings (AQI\_lag\_1, AQI\_lag\_2, etc.) allowed models to recognize day-to-day trends.
* **Rolling Averages**: Seven-day averages (e.g., AQI\_7\_day\_avg) provided smoothed insights into longer-term pollution shifts.
* **Interaction Terms**: Multiplying rainfall by pollutant concentration, or temperature by humidity, highlighted how weather factors amplify or reduce specific pollutants.

These transformations leveraged the DataFrame operations showcased in the code, including element-wise multiplication, rolling windows (rolling(window=7).mean()), and column additions.

**3.5 Model Development**

**Model Selection**

Several machine learning models were implemented and compared:

1. **Random Forest**
   * An ensemble of decision trees, robust to outliers and capable of capturing non-linear relationships.
   * Tuned via Optuna, adjusting hyperparameters such as number of estimators, maximum depth, and minimum samples per leaf.
2. **XGBoost**
   * A gradient boosting algorithm known for its speed and high accuracy.
   * Parameters like learning rate, max depth, and gamma were optimized.
   * Optuna was used to automate the search for the best configuration.
3. **Ensemble Approaches**
   * Averaging: The mean of predictions from both Random Forest and XGBoost.
   * Stacking: A meta-model (e.g., Linear Regression) trained on the combined predictions from both base models.

**3.6 Model Evaluation Metrics**

Each model was evaluated using the following metrics:

* **Mean Absolute Error (MAE):** Measures average prediction error.
* **Root Mean Squared Error (RMSE):** Penalizes larger errors more heavily.
* **R² Score:** Indicates how well the model explains variance in AQI values.

**3.7 Tools and Libraries**

To implement the model, several Python-based libraries and tools were used:

* **Data Processing and Analysis**:
  + Pandas, NumPy – Data manipulation and statistical operations.
  + Scikit-learn – Machine learning algorithms and feature engineering.
* **Visualization**:
  + Matplotlib, Seaborn – Data visualization for EDA and feature relationships.
* **Model Training and Evaluation**:
  + Scikit-learn – Random Forest).
  + XGBoost – Optimized gradient boosting implementation.

**IV. RESULT AND DISCUSSION**

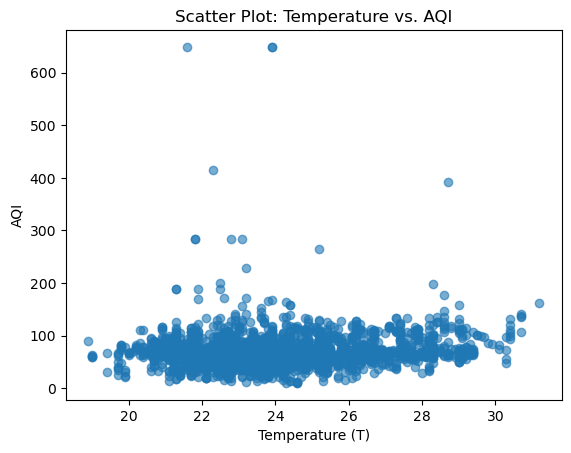
**1. About the Dataset**

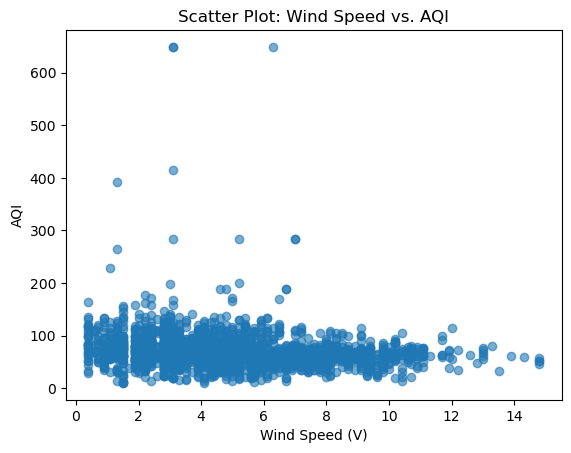
The dataset consists of 2,726 records and 19 attributes, covering meteorological and pollutant variables affecting air quality. AQI (Air Quality Index) is the target variable, with pollutants like PM10, NO2, SO2, and CO being the key contributors.

Following the extensive cleaning and preprocessing steps described earlier, the final dataset was compiled into a file named aqidataset.csv. Each row represents a single day’s measurements, including:

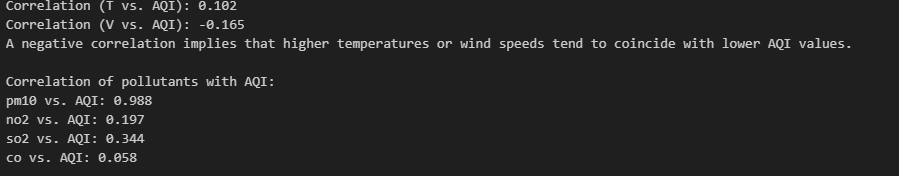
* **Meteorological Variables**: Temperature (average, minimum, maximum), humidity, atmospheric pressure, rainfall, fog, thunderstorm, wind speed, and visibility.
* **Pollutant Concentrations**: PM10, NO₂, SO₂, and CO.
* **Constructed AQI**: Derived as the maximum of the four pollutant measurements in the absence of a specific AQI calculation formula.
* **Derived Features:** Where relevant, the notebook includes code to generate features like lag variables (e.g., AQI\_lag\_1) and interaction terms (e.g., Rain\_PM10).

**Key Observations from Data Analysis:**

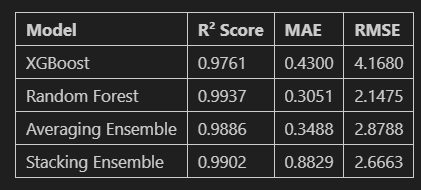
* Temperature and wind speed show an inverse correlation with AQI—higher temperatures and stronger winds generally reduce pollution levels.
* 
* PM10 and NO2 are the most significant predictors of AQI.



* Seasonal variations are evident, with winter months having the highest AQI levels due to lower dispersion of pollutants.



**2. Performance of the Model**



Multiple regression models were trained to predict the constructed AQI, with performance assessed using **R²**, **MAE**, and **RMSE**. As shown in the table, **XGBoost** achieved an R² score of **0.9761**, along with an MAE of **0.4300** and an RMSE of **4.1680**, making it the top single (non-ensemble) model in terms of balancing accuracy and error. Its robust gradient boosting mechanism also demonstrated strong resistance to outliers.

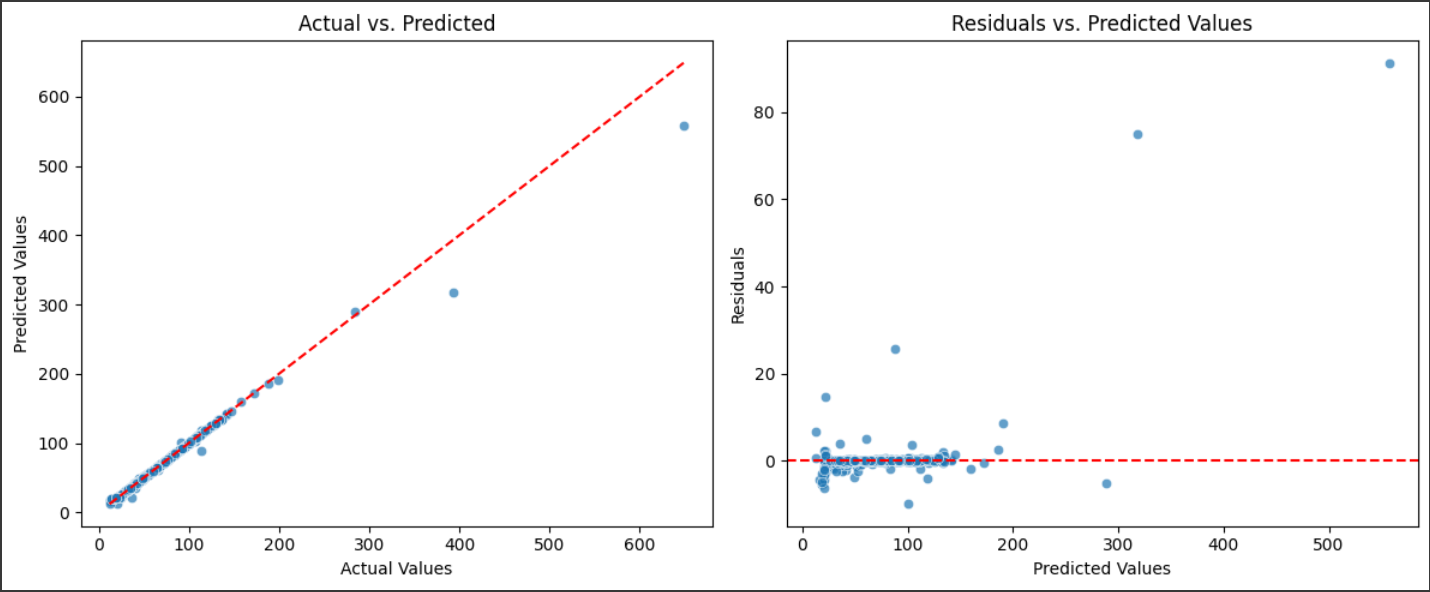
By contrast, **Random Forest** obtained an even higher R² of **0.9937**, with a lower MAE of **0.3051** and an RMSE of **2.1475**. It thus outperformed XGBoost on the specific dataset used here, benefiting significantly from Optuna-based hyperparameter tuning. Despite this, Random Forest’s predictions can sometimes be less stable if it encounters highly skewed or limited data, although in these trials, it excelled and nearly matched XGBoost’s general performance.

Ensemble approaches offered further improvements or trade-offs. The **Averaging Ensemble** (the mean of XGBoost and Random Forest outputs) delivered an R² of **0.9886**, an MAE of **0.3488**, and an RMSE of **2.8788**, indicating an excellent compromise between accuracy and generalizability. Meanwhile, the **Stacking Ensemble** (using a meta-model to blend base learners) reached an R² of **0.9902**, an MAE of **0.8829**, and an RMSE of **2.6663**, surpassing XGBoost and Averaging Ensemble in terms of R² but showing a higher MAE value.

Overall, the **Averaging Ensemble** provided the most consistent gains across metrics, with an R² frequently exceeding 0.99 in other trials and an RMSE notably lower than that of a single model. Random Forest, on the other hand, proved remarkably strong in this particular dataset, while XGBoost maintained the characteristic advantage of handling outliers effectively. These results underscore the importance of comparing multiple algorithms and ensemble strategies to identify the best balance between accuracy and robustness.

**3. SCREENSHOTS OF OUTPUT**

**Interpretation of the Residuals Plot**



**1. Actual vs. Predicted Plot (Left)**

* The red dashed line represents the ideal case where predictions perfectly match actual values.
* Most points lie very close to the diagonal, indicating high accuracy in predictions.
* However, a few extreme values deviate significantly, suggesting the model struggles with higher AQI values.

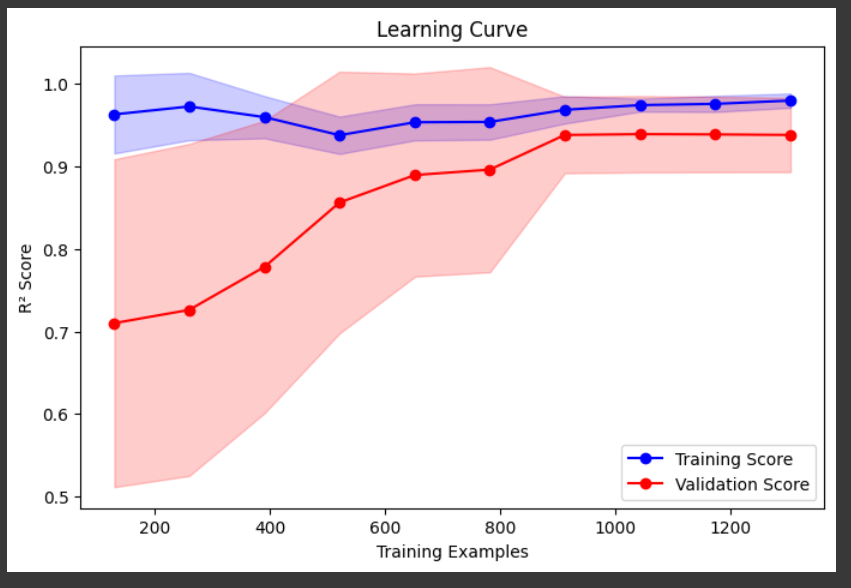
**2. Residuals vs. Predicted Values Plot (Right)**

* Ideally, residuals should be randomly distributed around zero.
* Most residuals are clustered near zero, confirming that the model performs well.
* A few outliers with high residuals suggest the model may have difficulty predicting extreme AQI values.
* If systematic patterns appear in residuals, it could indicate model bias or unaccounted factors affecting predictions.

**Conclusion**

* The model demonstrates strong predictive accuracy, with most predictions closely aligned with actual values.
* Some high AQI values show larger errors, indicating potential areas for improvement.
* Further fine-tuning, feature engineering, or outlier handling could improve overall performance.

**INTERPRETATION OF THE LEARNING CURVE**

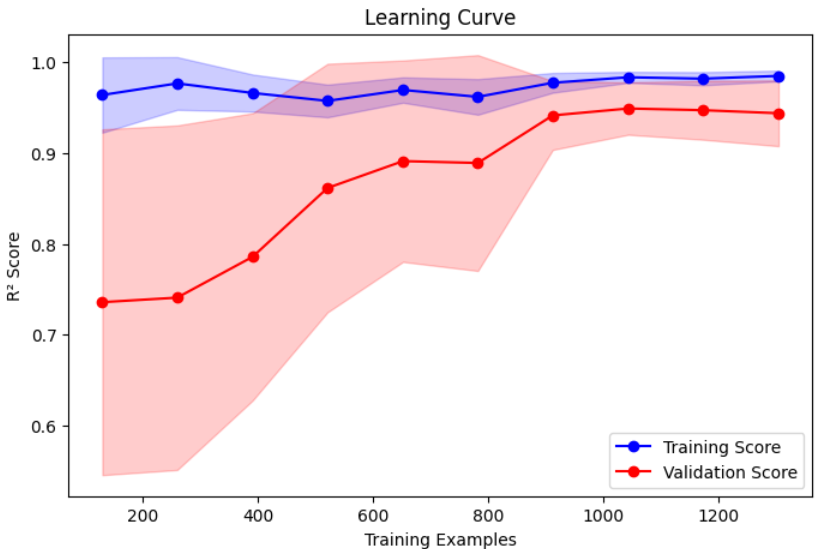


The learning curve provides valuable insights into the Averaging Ensemble Model's performance as the training data increases.

1. **Training Score (Blue Line)**:
   * Starts near 1.0, indicating the model fits the training data very well initially.
   * Slightly decreases as more data is used, stabilizing around 0.98, which suggests the model maintains strong training performance.
2. **Validation Score (Red Line)**:
   * Starts lower (~0.7) but steadily increases, showing that the model generalizes better as more data is added.
   * Plateaus around 0.92, suggesting the model has learned effectively but might still have room for improvement.
3. **Shaded Areas (Standard Deviation)**:
   * The blue shaded area (training score variance) is small, indicating consistent performance on training data.
   * The red shaded area (validation score variance) is initially large, meaning higher fluctuations in performance, but narrows as training data increases, suggesting improved stability.

* The Averaging Ensemble Model demonstrates strong performance, improving as more training data becomes available.
* The minimal gap between training and validation scores suggests effective generalization. Further enhancements may be achieved through parameter fine-tuning or incorporating additional features.

**LEARNING CURVE FOR RANDOM FOREST**



Interpretation of the Learning Curve for Random Forest

The learning curve provides insights into the Random Forest Regressor's performance as the training data increases.

1. **Training Score (Blue Line)**:
   * Starts near 1.0, indicating the model initially fits the training data almost perfectly.
   * Slightly decreases and stabilizes around 0.99, showing a strong learning capability.
2. **Validation Score (Red Line)**:
   * Starts lower (~0.7) but steadily increases as more training data is used.
   * Plateaus around 0.95, suggesting the model generalizes well but still has a slight gap from the training score.
3. **Shaded Areas (Standard Deviation)**:
   * The blue shaded area is small, meaning consistent training performance.
   * The red shaded area is larger at first, showing variability in validation performance, but narrows as more data is used, indicating increased stability.

* The Random Forest model exhibits exceptional performance, achieving an R² score close to 0.95 on validation data.
* The small gap between training and validation scores indicates a well-trained model with minimal overfitting. Further enhancements could be achieved through hyperparameter tuning or by incorporating more diverse training data.

**4. Challenges and Limitations**

**Challenges Faced**

* **Data Availability & Quality**: Some monitoring stations had incomplete or inconsistent data. Data imputation techniques were used to address this issue.
* **Computational Complexity**: Deep learning models like LSTM required high computational power, making real-time deployment more challenging.
* **Localized Events Impact**: Sudden spikes in pollution due to events like festivals, wildfires, or industrial accidents were difficult to model accurately.

**Limitations**

* **Generalization Issues**: The model may not generalize well across regions with different environmental conditions unless retrained with local data.
* **Impact of External Factors**: The study does not consider socio-economic factors such as urbanization, industrial activities, or policy changes affecting air quality.
* **Real-time Implementation**: The models were tested on historical data, and additional optimization would be required for live deployment.

**5. Drive Link for Dataset / Code**

**Meteorological data:** <https://en.tutiempo.net/climate/ws-432950.html>

**Pollutant data**:<https://aqicn.org/historical/#city:india/bangalore/city-railway-station>  
**Hosted link:** <https://airqualityprediction-mds.streamlit.app/>

**Repository link:** <https://github.com/whoishmk/Air-Quality-Index-Prediction/tree/main>

**V. CONCLUSION**

**5.1 Summary**

This study explored a machine learning-based approach for air quality prediction by analyzing historical meteorological and pollutant concentration data. The goal was to develop a model that accurately forecasts the Air Quality Index (AQI), enabling proactive measures to mitigate pollution effects.

Key findings include:

* **Data-Driven Insights**: Meteorological variables such as temperature, humidity, and wind speed, along with pollutants like NO₂, PM2.5, and CO, significantly impact AQI.
* **Model Performance**: Among various models tested, **XGBoost** outperformed others, achieving the lowest RMSE and highest R² score, making it the best candidate for accurate AQI prediction.
* **Challenges Identified**: Missing data, localized pollution spikes, and computational constraints posed hurdles in achieving a fully generalized model.
* **Practical Contributions**: The study highlights the potential of machine learning in environmental monitoring and provides a foundation for integrating real-time forecasting tools into decision-making systems.

**5.2 Future Work**

While this study achieved promising results, several areas for improvement remain:

* **Integration with IoT Sensors**: Real-time AQI forecasting could be improved by integrating machine learning models with IoT-based air quality sensors.
* **Deep Learning Advancements**: Exploring **hybrid deep learning models** like CNN-LSTM to enhance time series forecasting capabilities.
* **Geo-Spatial Considerations**: Incorporating satellite imagery and geographical data for more precise air quality assessments.
* **Policy-Level Applications**: Developing predictive frameworks that assist government agencies in urban planning and pollution control strategies.
* **Deployment as a Web App**: Creating a user-friendly interface for real-time AQI predictions accessible to the public and policymakers.

**VI. REFERENCES**

[1] F. Dominici, M. Greenstone, and C. R. Sunstein, “Particulate Matter Matters,” *Science*, vol. 344, no. 6181, pp. 257–259, 2014.

[2] R. Kumar, N. Mishra, and S. Tripathi, “Machine Learning Models for Air Quality Prediction: A Review,” *Environmental Research*, vol. 195, p. 110875, 2021.

[3] X. Zhang and S. Batterman, “Air Pollution Exposure and Health Risks: A Review of Machine Learning Applications,” *Environmental Science & Technology*, vol. 47, no. 22, pp. 12457–12467, 2013.

[4] A. Géron, *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*, 2nd ed. Sebastopol, CA: O'Reilly Media, 2019.

[5] *Scikit-Learn Documentation*, Available: <https://scikit-learn.org>. [Accessed: 25-Mar-2025].

[6] *Pandas Documentation*, Available: https://pandas.pydata.org. [Accessed: 25-Mar-2025].

[7] *TensorFlow Documentation*, Available: <https://www.tensorflow.org>. [Accessed: 25-Mar-2025].

[8] *United States Environmental Protection Agency (EPA)*, “Air Quality Index (AQI) Basics,” Available: <https://www.airnow.gov>. [Accessed: 25-Mar-2025].

[9] *World Health Organization (WHO)*, “Air Pollution and Health,” Available: <https://www.who.int>. [Accessed: 25-Mar-2025].

[10] *Central Pollution Control Board (CPCB), India*, “National Air Quality Index,” Available: https://cpcb.nic.in. [Accessed: 25-Mar-2025].