IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 7, June 2026
Impact Factor: 8.2

# Air Pollution Prediction Using Machine Learning for Air Quality Index (AQI)

**Zeba Mulla[1] and Kavita Koli[2]**
Research Scholar, Department of Computer Scienc
, Shivaji University, Kolhapur, Maharashtra, India.

**Abstract:** *Air pollution is a serious environmental problem that affects human health. Predicting air quality in advance can help people take preventive actions. In this research, a machine learning based system is developed to predict the Air Quality Index (AQI). The model uses pollution parameters such as PM2.7, PM10, NO2, SO2, CO, and O3. Multiple machine learning algorithms such as Linear Regression, Decision Tree, Random Forest, Support Vector Machine, and K-Nearest Neighbors are used and compared. The best performing model is selected based on accuracy. The system can predict AQI and classify air quality levels such as Good, Moderate, Poor, Very Poor, and Severe.*

**Keywords:** Air Pollution, Air Quality Index (AQI), Machine Learning, AQI Prediction, Environmental Monitoring, Random Forest, Linear Regression, Data Analysis.

## I. INTRODUCTION

Air pollution is a serious issue affecting cities around the world. High pollution levels can cause health problems such as respiratory diseases, heart problems, and environmental damage.

The **Air Quality Index (AQI)** is used to measure pollution levels. AQI values help people understand whether the air quality is safe or harmful.

Traditional systems only show the current air quality but do not predict future pollution levels. Machine learning can help solve this problem by analyzing historical pollution data and predicting AQI.

This research focuses on building a machine learning model that predicts AQI based on different pollutant levels.

**Dataset Description:**

The dataset used in this project contains pollution parameters such as:

- PM2.7
- PM10
- NO□
- SO□
- CO
- O□
- AQI (target value)

These parameters are used as **input features**, while AQI is the **output value to be predicted**.

## II. LITERATURE REVIEW

- Machine Learning improves AQI prediction performance.
- Common algorithms used: Linear Regression, Decision Tree, Random Forest, SVM.
- Random Forest gives better accuracy for complex data.
- Main paramTraditional methods like regression were used earlier but had low accuracy.
- eters used: PM2.7, PM10, NO□, SO□, CO, O□.
- Large datasets help in better prediction results.
- Models are evaluated using MSE and R² score.


DOI: 10.48175/IJARSCT-36887

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 7, June 2026
Impact Factor: 8.2

- Existing studies have limitations like small data and ignoring weather factors.
- There is a need for more accurate and real-time AQI prediction models.

## III. RESEARCH METHODOLOGY

The following steps were used in the project:

**1. Data Collection**

Air pollution data was collected and stored in a dataset.

**2. Data Preprocessing**

The dataset was cleaned and prepared for training.

**3. Feature Selection**

The pollution parameters were used as input variables.

**4. Data Splitting**

The dataset was divided into:

- **Training data (80%)**
- **Testing data (20%)**

**5. Feature Scaling**

StandardScaler was used to normalize the data.

**6. Model Training**

- Five machine learning algorithms were used:
- Linear Regression
- Decision Tree
- Random Forest
- Support Vector Machine
- K-Nearest Neighbors

**7. Model Evaluation**

Models were evaluated using:

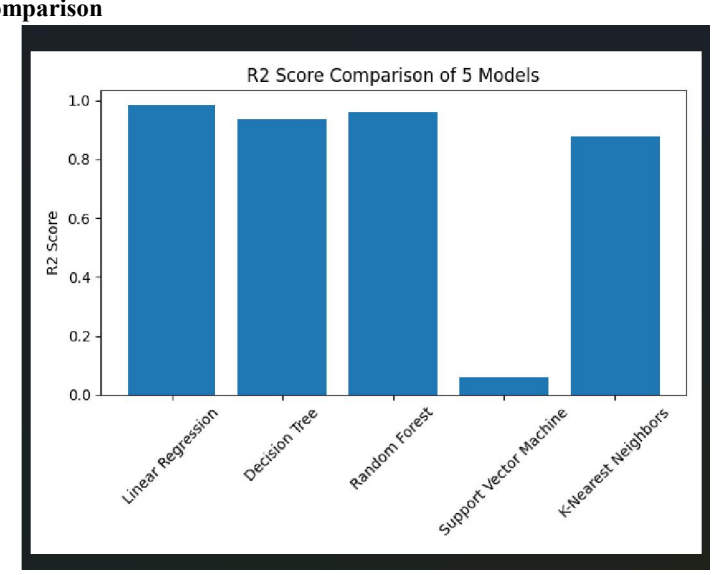
- Mean Squared Error (MSE)
- $R^2$ Score

## IV. RESULTS AND GRAPH EXPLANATION

The $R^2$ score measures how well the model predicts AQI values. A higher RRscore means better prediction accuracy. From the graph:

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 7, June 2026
Impact Factor: 8.2

**Graph 1: $R^2$ Score Comparison**

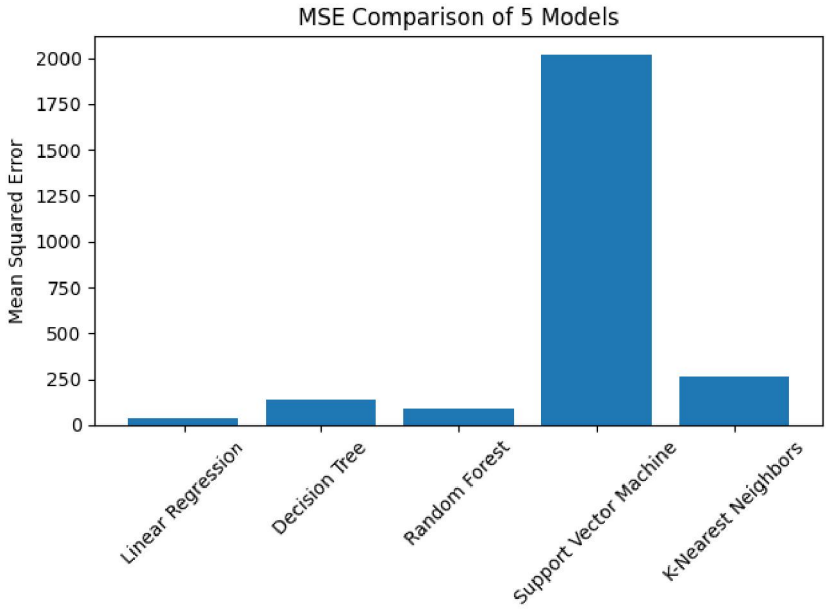



**Explanation :**

Linear Regression gives the **highest accuracy (0.984)**.
Random Forest also performs well.
Support Vector Machine shows very low performance.
Therefore, Linear Regression is selected as the **best model**.

**Graph 2: Mean Squared Error Comparison:**

Mean Squared Error (MSE) measures the difference between predicted AQI and actual AQI values. Lower MSE means better prediction performance.



IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 7, June 2026
Impact Factor: 8.2

**Explanation :**

- Linear Regression has the lowest error (34).
- Random Forest also shows good performance.
- Support Vector Machine has the highest error, meaning poor prediction.
- This confirms that Linear Regression is the best model for this dataset.

**AQI Prediction System:**

The developed system allows users to enter pollution values:

PM2.5, PM10, $NO_2$, $SO_2$, CO, $O_3$

The trained model predicts the AQI value and classifies it into categories:

| AQI Range | Category: |
|---|---|
| 0 – 70 | Good |
| 71 – 100 | Moderate |
| 101 – 200 | Poor |
| 201 – 300 | Very Poor |
| 300+ | Severe |

This helps users understand air quality levels easily.

**Conclusion:**

This research developed a machine learning system to predict Air Quality Index using pollution parameters. Multiple algorithms were tested and compared.

The results show that Linear Regression achieved the best performance with an $R^2$ score of 0.984 and the lowest error.

**Future Work:**

Future improvements may include:

- Using real-time pollution data
- Applying deep learning models
- Integrating weather data
- Developing a web or mobile application
- Expanding the dataset for better accuracy

**REFERENCES**

1. World Health Organization, “Air Pollution and Health,” 2018.
2. Central Pollution Control Board, “National Air Quality Index,” Government of India, 2014.
3. UCI Machine Learning Repository, “Air Quality Data Set,” Available online.
4. L. Breiman, “Random Forests,” Machine Learning Journal, 2001.
5. C. Cortes and V. Vapnik, “Support-Vector Networks,” Machine Learning Journal, 1995.
6. Scikit-learn Documentation, Available: https://scikit-learn.org
7. Pandas Documentation, Available: https://pandas.pydata.org
8. NumPy Documentation, Available: https://numpy.org