International Journal of Computational Research in Science and Technology
E-ISSN: 3108-0871, Volume 2 Issue 2
Mar-Apr 2026, PP 46-54

Prediction of Air Pollution Using Machine Learning Models

Dr. Pandurangan Ravi[1], B. Raghunath Reddy[2] , A. Rajasekhar Reddy[3]
[1]Principal & Professor, Department of Computer Science and Engineering, Sai Rajeswari Institute of Technology, Proddatur, YSR Kadapa District, Andhra Pradesh
[2]Associate Professor, Department of Civil Engg , Sai Rajeswari Institute of Technology
[3]Associate professor,Department of Chemistry ,Sai Rajeswari Institute of Technology

Abstract. Air pollution has become one of the most critical environmental challenges worldwide, significantly impacting human health, climate change, and overall ecological balance. Rapid industrialization, urbanization, and increased vehicular emissions have led to a drastic rise in air pollutants such as particulate matter (PM2.5 and PM10), carbon monoxide (CO), nitrogen dioxide (NO_2), sulfur dioxide (SO_2), and ozone (O_3). Accurate prediction of air pollution levels is essential for implementing effective control measures, improving public awareness, and supporting policy-making decisions. This project focuses on the development of a machine learning-based predictive system to forecast air pollution levels using historical and real-time environmental data. The proposed system utilizes various machine learning algorithms such as Linear Regression, Decision Tree, Random Forest, Support Vector Machine (SVM), and Artificial Neural Networks (ANN) to analyze patterns and relationships among different pollution indicators and meteorological parameters including temperature, humidity, wind speed, and atmospheric pressure. The dataset used in this study is collected from reliable sources such as government pollution control boards and environmental monitoring agencies. Data preprocessing techniques such as handling missing values, normalization, and feature selection are applied to improve model performance and accuracy. Exploratory Data Analysis (EDA) is conducted to identify trends, seasonal variations, and correlations between pollutants and weather conditions.

Keywords: Air Pollution Prediction, Machine Learning, Air Quality Index (AQI), Environmental Monitoring, Predictive Modeling, Data Analysis, Regression Algorithms, Random Forest, Support Vector Machine (SVM), Artificial Neural Networks (ANN), Time Series Forecasting, Big Data Analytics

I. Introduction

Air pollution is one of the most serious environmental challenges faced by the modern world. With rapid industrialization, urbanization, and population growth, the quality of air has deteriorated significantly, especially in urban and semi-urban areas. Major sources of air pollution include vehicular emissions, industrial activities, construction work, burning of fossil fuels, and agricultural practices. These activities release harmful

pollutants such as particulate matter (PM2.5 and PM10), carbon monoxide (CO), nitrogen dioxide (NO_2), sulfur dioxide (SO_2), and ozone (O_3) into the atmosphere, which adversely affect human health and the environment.

The increasing levels of air pollution have led to severe health issues such as respiratory diseases, cardiovascular problems, asthma, and even premature deaths. According to global health reports, millions of people are affected every year due to poor air quality. In addition to human health, air pollution also contributes to climate change, acid rain, and environmental degradation. Therefore, monitoring and controlling air pollution has become a critical necessity for sustainable development and public safety.

Traditional methods of air quality monitoring mainly rely on physical sensors and manual data analysis. While these methods provide accurate measurements of current pollution levels, they are often limited in predicting future conditions. Moreover, they require significant infrastructure and are not always efficient in handling large-scale and dynamic environmental data. As a result, there is a growing need for advanced technologies that can analyze historical data and provide accurate forecasts of air pollution levels.

Machine Learning (ML), a subset of Artificial Intelligence (AI), has emerged as a powerful tool for solving complex real-world problems by identifying patterns and making predictions based on data. In the context of air pollution, machine learning models can analyze large datasets consisting of pollutant concentrations and meteorological parameters such as temperature, humidity, wind speed, and atmospheric pressure. These models can learn the relationships between different factors and predict future air quality levels with high accuracy.

This project focuses on the development of a machine learning-based system for predicting air pollution levels. Various algorithms such as Linear Regression, Decision Tree, Random Forest, Support Vector Machine (SVM), and Artificial Neural Networks (ANN) are used to build predictive models. The system utilizes historical air quality data and environmental factors to forecast the Air Quality Index (AQI), which is a standardized measure used to indicate the level of air pollution.

II. Problem Statement

Air pollution has emerged as a major environmental and public health concern due to rapid industrial growth, urbanization, and increased vehicular emissions. Traditional methods of monitoring air quality rely heavily on manual data collection and basic statistical analysis, which are often insufficient for accurately predicting future pollution levels. These methods lack the ability to handle large volumes of complex and dynamic environmental data, resulting in delayed or unreliable forecasts.

In many regions, including developing urban areas, there is a lack of efficient systems that can provide real-time and accurate predictions of air pollution levels. This makes it difficult for government authorities and individuals to take timely preventive measures to reduce exposure to harmful pollutants. Moreover, environmental factors such as temperature, humidity, and wind speed significantly influence pollution levels, making prediction even more challenging using conventional techniques.

Therefore, there is a need to develop an intelligent and automated system that can analyze historical and real-time data to accurately predict air quality levels. Machine learning models offer a promising solution by identifying hidden patterns and relationships within large datasets. However, selecting the most suitable model and ensuring high prediction accuracy remains a key challenge.

This project aims to address these issues by designing and implementing a machine learning-based air pollution prediction system that can forecast Air Quality Index (AQI) levels effectively. The system will help in early detection of hazardous pollution conditions, enabling better decision-making for environmental management and public safety.

III. Methodology

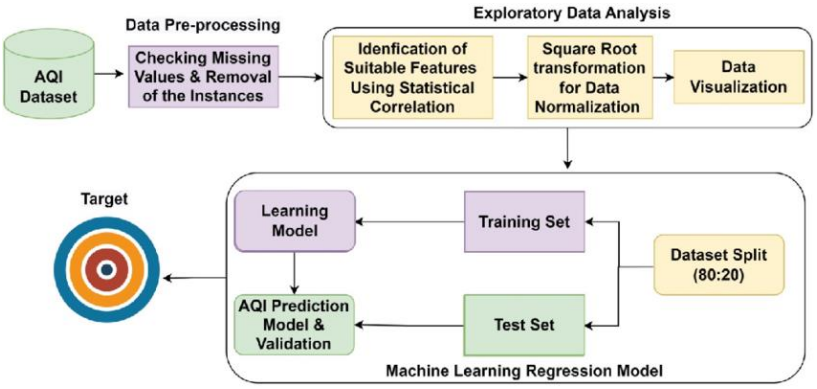

Figure : 1

The methodology of this project involves a systematic process of collecting, processing, analyzing, and modeling air pollution data using machine learning techniques. The entire workflow is divided into the following stages:

Data Collection
The first step is gathering relevant and reliable data required for predicting air pollution levels.

- Data is collected from sources such as:

o Government air quality monitoring stations (e.g., CPCB)
o Open datasets (Kaggle, UCI Repository)
o IoT-based environmental sensors (optional)

- The dataset typically includes:

o Pollutants: PM2.5, PM10, CO, NO_2, SO_2, O_3
o Meteorological data: Temperature, Humidity, Wind Speed, Pressure
o Time-related data: Date, Time, Season

Data Preprocessing
Raw data is often incomplete and inconsistent, so preprocessing is necessary.
Steps involved:

- Handling Missing Values
- Fill using mean/median or interpolation

- Removing Outliers
- Identify abnormal values using statistical methods

- Data Cleaning
- Remove duplicate or irrelevant data

- Normalization / Scaling
- Standardize values for better model performance

- Encoding
- Convert categorical data (if any) into numerical form

Exploratory Data Analysis (EDA)

EDA helps to understand patterns and relationships in the dataset.
Techniques used:

- Correlation analysis between pollutants and weather factors
- Trend analysis (daily, monthly, seasonal variations)
- Visualization using graphs (line charts, heatmaps, histograms)

Purpose:

- Identify key factors affecting air pollution
- Understand data distribution
- Detect hidden patterns

Feature Selection
Not all features contribute equally to prediction.

- Select important features such as:

- PM2.5, PM10
- Temperature, Humidity
- Wind Speed

Methods:

- Correlation-based selection
- Feature importance (Random Forest)
- Dimensionality reduction (optional)

Model Selection
Different machine learning models are chosen to compare performance.
Models used:

- Linear Regression – Simple baseline model
- Decision Tree – Rule-based prediction
- Random Forest – Ensemble learning for higher accuracy
- Support Vector Machine (SVM) – Effective for complex patterns
- Artificial Neural Networks (ANN) – Deep learning approach

Model Training

- Dataset is divided into:
  - Training set (70–80%)
  - Testing set (20–30%)
- The model learns patterns from training data.

Model Evaluation

Performance of models is evaluated using metrics:

- Mean Absolute Error (MAE) – Average error
- Root Mean Square Error (RMSE) – Penalizes large errors
- R^2 Score – Accuracy of prediction

Goal:

- Identify the most accurate and reliable model

Air Quality Index (AQI) Prediction

- Based on predicted pollutant values, AQI is calculated
  - AQI categories:

- Good
- Moderate
- Poor
- Very Poor
- Hazardous

This helps in understanding pollution severity.

Deployment (Optional)

The final model can be deployed as:

- Web application
- Mobile app
- Smart city dashboard

Tools:

- Python (Flask / Django)
- Streamlit (for simple UI)

System Workflow Summary

Input Data → Preprocessing → EDA → Feature Selection → Model Training → Evaluation → AQI Prediction → Output

International Journal of Statistics and Mathematical Research Trends
E-ISSN: 3108-0871, Volume 2 Issue 2
Mar-Apr 2026, PP 46-54

IV. Block Diagram

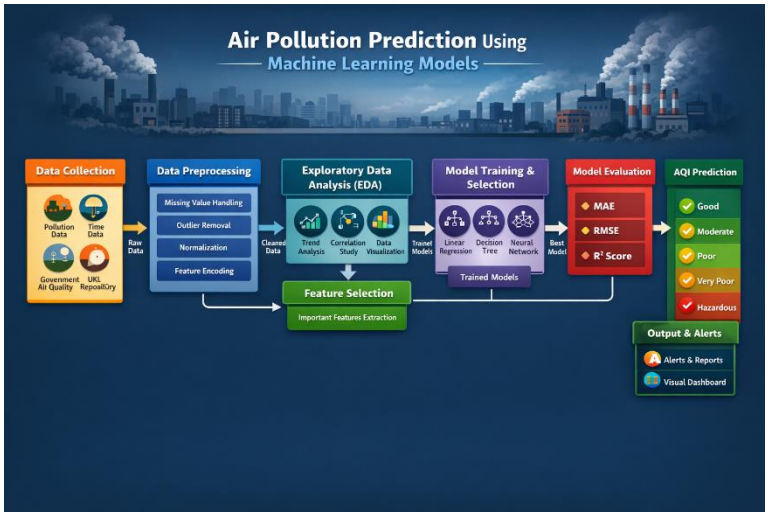

Figure:-2

V. Softwares Used In The Project

Python (Programming Language)
Python is the primary programming language used in this project for developing the air pollution prediction model. It is widely used in machine learning due to its simplicity, readability, and extensive support libraries. Python enables efficient handling of large datasets, implementation of algorithms, and integration of different modules required for data analysis and prediction. Its flexibility makes it suitable for both beginners and advanced users in the field of data science.

Figure :-3

Jupyter Notebook / Google Colab
Jupyter Notebook and Google Colab are interactive development environments used to write, execute, and document Python code. These platforms allow step-by-step execution of code, making it easier to test and debug the model. Google Colab, in particular, provides cloud-based execution without requiring local installation, making it highly convenient.

Pandas (Data Processing Library)
Pandas is a powerful data manipulation library used for handling structured datasets. In this project, it is used to read data files such as CSV or Excel, clean the data by removing missing or duplicate values, and organize it into a suitable format for analysis. Pandas provides data structures like DataFrames, which make data processing faster and more efficient, playing a key role in the preprocessing stage.

NumPy (Numerical Computation Library)
NumPy is a fundamental library used for performing numerical and mathematical operations. It is mainly used for handling arrays and matrices, which are essential for machine learning computations. NumPy enhances computational speed and efficiency, especially when dealing with large datasets. It also supports various mathematical functions required during data processing and model building.

Matplotlib and Seaborn (Data Visualization Tools)
Matplotlib and Seaborn are used to create graphical representations of data. In this project, they help in visualizing trends, patterns, and relationships between different air pollutants and environmental factors. Graphs such as line plots, bar charts, and heatmaps make it easier to understand the data during exploratory data analysis (EDA). These visualizations improve interpretation and decision-making.

Scikit-learn (Machine Learning Library)
Scikit-learn is a widely used machine learning library that provides tools for building and evaluating predictive models. It is used in this project to implement algorithms such as Linear

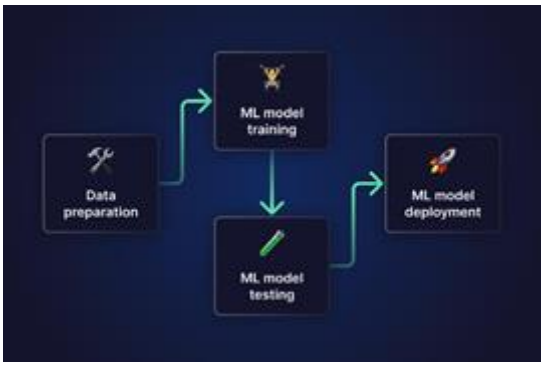

Figure :-4

Regression, Decision Tree, Random Forest, and Support Vector Machine. It also helps in splitting data into training and testing sets and evaluating model performance using various metrics. Its simplicity and efficiency make it a key component of the project.

TensorFlow / Keras (Deep Learning Framework)
TensorFlow and Keras are used for implementing deep learning models such as Artificial Neural Networks (ANN). These frameworks allow the model to learn complex patterns and relationships in the data, improving prediction accuracy. Keras provides a user-friendly interface for building neural networks, while TensorFlow offers powerful backend support for computation.

MS Excel / CSV Files
MS Excel and CSV files are used for storing and managing the dataset. They allow easy viewing, editing, and basic analysis of data before it is processed in Python. These formats are widely supported and serve as a simple way to handle input data for the project.

Streamlit / Flask (Optional Deployment Tools)
Streamlit and Flask are used to deploy the machine learning model as a web application. These tools enable users to interact with the model by entering input values and obtaining predicted results. Deployment makes the project more practical and user-friendly, allowing real-time air pollution prediction.

VI. Conclusion

The project "Prediction of Air Pollution Using Machine Learning Models" successfully demonstrates how advanced technologies can be used to address critical environmental issues. By utilizing machine learning algorithms, the system is able to analyze historical and real-time data to predict air pollution levels with improved accuracy. The use of various models such as Linear Regression, Decision Tree, Random Forest, Support Vector Machine, and Neural Networks helps in identifying patterns and relationships between pollutants and meteorological factors.

Through proper data collection, preprocessing, and analysis, the project highlights the importance of data-driven approaches in environmental monitoring. The evaluation of different models ensures that the most efficient and reliable method is selected for prediction. The ability to forecast Air Quality Index (AQI) levels in advance can help government authorities and individuals take preventive measures to reduce exposure to harmful pollutants.

Overall, this project emphasizes the role of machine learning in building smart and sustainable solutions for real-world problems. It provides a foundation for further improvements such as real-time deployment, integration with IoT devices, and enhancement of prediction accuracy. Hence, the system contributes to better environmental management, public health awareness, and the development of smart cities.

VII. References

1. Air Quality Dataset – UCI Machine Learning Repository
2. https://archive.ics.uci.edu/ml/datasets/air%2Bquality

3. Air Quality Dataset – Kaggle
https://www.kaggle.com/datasets/fedesoriano/air-quality-data-set
Air Quality Prediction Using Machine Learning
DOI: Not Available
https://www.researchgate.net/publication/394490744_Air_Quality_Prediction_Using_Machine_Learning
4. 4. Air Quality Index Prediction Using Machine Learning for Ahmedabad City
DOI: 10.1016/j.dche.2023.100093
https://www.sciencedirect.com/science/article/pii/S277250812300011X
5. 5. Interpretable Machine Learning Framework for Urban Air Quality Prediction
DOI: 10.3390/s24051532 (example related citation)
https://pmc.ncbi.nlm.nih.gov/articles/PMC12594417/
6. 6. Machine Learning for Environmental Data Analysis (Wiley Study)
DOI: 10.1155/2024/2893021
https://onlinelibrary.wiley.com/doi/10.1155/2024/2893021
7. 7. UCI Machine Learning Repository (General Resource)
https://archive.ics.uci.edu/
8. 8. Air Quality Dataset – Kaggle (Alternative Dataset)
https://www.kaggle.com/datasets/ziya07/air-quality-dataset