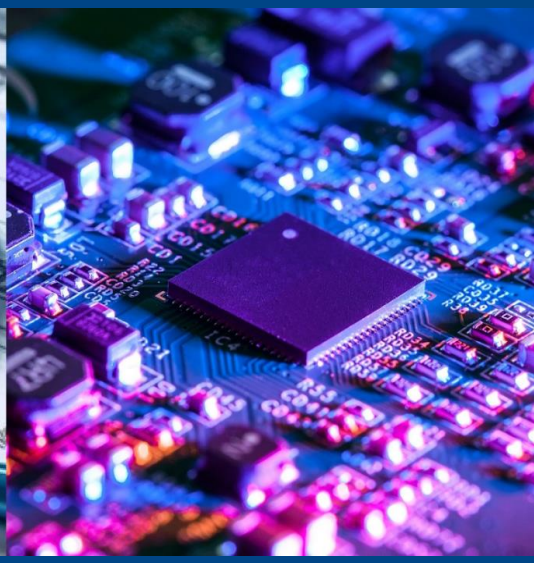
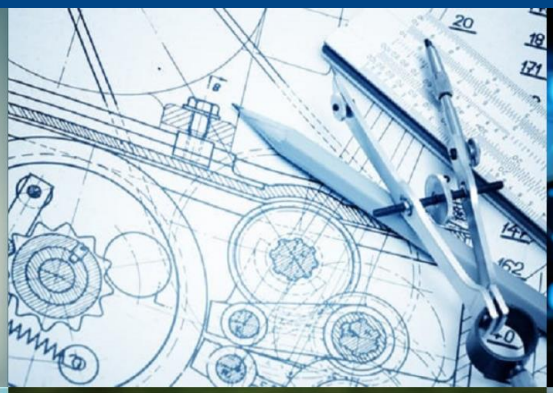
ISSN: 2582-7219

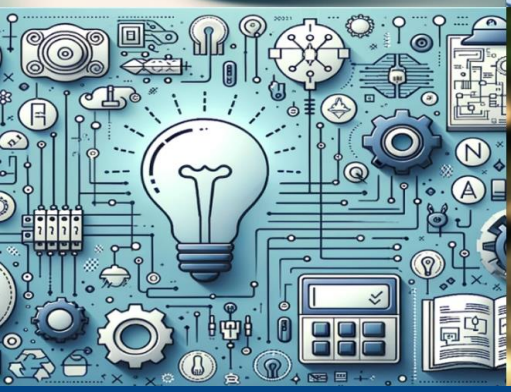




International Journal of Multidisciplinary Research in Science, Engineering and Technology

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.206 **Volume 8, Issue 5, May 2025**

© 2025 IJMRSET | Volume 8, Issue 5, May 2025| DOI:10.15680/IJMRSET.2025.0805273

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Air Quality index Prediction using Machine Learning

Digambar Dashrath Wagh

Postgraduate Student, Dept. of Master of Computer Application, APCOER, Pune, India

Prof. Purvesh Wagh

Asst. Professor, Dept. Of Master of Computer Application, APCOER, Pune, India

ABSTRACT: The rapid evolution of artificial intelligence (AI) and machine learning (ML) technologies has unlocked innovative approaches for analyzing environmental datasets to derive predictive insights. A key area where this innovation is being applied is in forecasting the Air Quality Index (AQI), which plays a vital role in safeguarding public health and shaping urban environmental strategies. This study investigates methods to improve the accuracy of AQI forecasts by integrating several machines learning methodologies, focusing on three critical components: data preprocessing, feature selection, and model optimization. Data preprocessing involves transforming unstructured air quality data into a clean, organized format suitable for machine learning applications, ensuring consistency and reliability. Feature selection helps identify the most impactful pollutants—such as PM2.5, NO, SO_2, and CO—that heavily influence AQI fluctuations. Model optimization refers to fine-tuning machine learning algorithms, including Random Forest, Support Vector Machines, and Gradient Boosting, to enhance predictive accuracy. Through a series of experiments and performance comparisons, this research illustrates how these combined strategies significantly improve AQI prediction performance. The outcomes contribute to the development of intelligent, data-driven environmental monitoring systems designed to support informed public health decisions and promote sustainable urban living.

KEYWORDS: Large Language Models (LLMs), Fine-Tuning, Prompt Design, Context Awareness, Code Generation

I. INTRODUCTION

Air pollution remains a big issue, affecting people's health and the balance of nature. The Air Quality Index (AQI) is a simple way to show how polluted the air is and to help people stay safe. Traditional AQI monitoring systems rely heavily on hardware sensors and static reports, which often fail to offer the predictive insights needed for proactive decision-making. With advancements in machine learning, it is now possible to develop models that anticipate AQI trends by analyzing patterns in environmental data.

This research focuses on advancing AQI prediction through machine learning by targeting three core areas:

- **Data preprocessing:** Cleaning and organizing raw environmental datasets to optimize their usefulness in model training.
- **Feature selection:** Identifying the pollutants and variables that most significantly influence AQI variations.
- **Model tuning**: means fine-tuning machine learning methods to boost prediction accuracy and make results more reliable

By combining these strategies, the study proposes a scalable and robust framework for AQI forecasting, aimed at supporting real-time pollution assessment and facilitating data-informed policy development.

Fine-Tuning in Machine Learning for Predicting Air Quality Index

In Air Quality Index (AQI) prediction, fine-tuning plays a key role in improving model performance by customizing pre-trained models with specific environmental data. Instead of training from scratch, the model is adjusted using smaller, relevant datasets that include local pollution levels and weather conditions. This approach helps the model become more accurate in identifying changes in air quality across different locations and time periods

Traditional approaches include:

© 2025 IJMRSET | Volume 8, Issue 5, May 2025| DOI:10.15680/IJMRSET.2025.0805273

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

- Supervised Fine- These uses labeled AQI data, often gathered from monitoring stations, to adjust the model for more accurate predictions.
- Transfer Learning- In this method, a model trained on general environmental patterns is adapted to focus specifically on AQI-related data.
- This aligns the model to the specific characteristics of the target area, such as local emission sources or seasonal trends.

Newer approaches like **efficient fine-tuning** techniques—such as low-rank adaptation (LoRA) and adapter modules—help reduce the amount of computing power needed, while still keeping prediction quality high. These methods are especially helpful in devices with limited resources, like mobile or edge devices used for real-time air quality monitoring.

II. PROMPT DESIGN IN CODE GENERATION FOR AQI PREDICTION SYSTEM

- Prompt design plays a vital role in generating accurate and useful code using large language models (LLMs), especially when working on technical projects like Air Quality Index (AQI) prediction. It involves crafting precise and well-structured input to guide the model in producing code that aligns with specific goals, such as building ML pipelines, data preprocessing scripts, or model evaluation tools.
- Key elements of effective prompt design include:
- Clarity Instructions: Prompts must clearly state what needs to be done. Vague prompts often result in incomplete or irrelevant code. Being direct about the task (e.g., "Write a Python function to preprocess AQI dataset") improves results.
- Relevant Examples: Including examples like sample data, code snippets, or expected output helps the model understand the task better and mimic the desired format or logic.
- Task Context: Providing background details—such as which libraries to use, the target ML algorithm, or performance metrics—helps generate code that fits the problem domain.
- Prompt design is especially useful when using few-shot or zero- shot approaches, where the model has little to no prior example. A well-crafted prompt can guide the model to understand patterns and generate useful code on its own.
- Additionally, organizing information within the prompt—putting important details up front and keeping a clean, structured format— can make a noticeable difference
- Even small changes in phrasing or order can change the outcome, which makes prompt writing not just a technical step, but a creative and thoughtful process in AI-assisted code generation.

III. CONTEXT AWARENESS IN CODE GENERATION FOR AQI PREDICTION

Context awareness is a key factor in generating accurate and meaningful code using large language models (LLMs), especially when developing systems for Air Quality Index (AQI) prediction. It involves understanding the specific background, goals, and structure of the task so the generated code matches the real-world needs of the project.

Key benefits of context-aware code generation include:

- Relevant Output: The model generates code that fits the specific task, such as handling missing pollution data or converting raw sensor inputs into usable features.
- Consistency with Project Goals: Context-aware prompts help produce code that follows the structure of the overall project—like integrating prediction models into dashboards or APIs.
- Smarter Defaults: LLMs can select more appropriate libraries or techniques (e.g., using scikit-learn for classification or matplotlib for AQI visualizations) when they understand the environment they are working in.
- By including context such as data format, desired outcome, or even previous code snippets in the prompt, developers can significantly improve the quality and usefulness of the generated code. In AQI prediction projects, this leads to faster development, fewer errors, and more practical machine learning solutions.

© 2025 IJMRSET | Volume 8, Issue 5, May 2025| DOI:10.15680/IJMRSET.2025.0805273

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

IV. EMPIRICAL ANALYSIS AND CASE STUDIES

Empirical analysis plays a pivotal part in understanding how machine knowledge models perform when applied to real-world Air Quality Index(AQI) prophecy tasks. By experimenting with factual datasets and assessing model issues, researchers and formulators can identify which algorithms work stylish under different environmental and data conditions.

Evaluation Matrix

To measure how well machine literacy models prognosticate Air Quality Index, we use the following crucial evaluation points

Mean Absolute Error (MAE) Tells us the average size of the difference between prognosticated and factual AQI values. Lower scores mean more accurate results.

Root Mean Squared Error (RMSE) Focuses further on bigger miscalculations in vaticination, making it easier to spot major crimes.

R^2 Score (R- squared) Shows how well the model captures patterns in the data. A score near to 1 means the model fits the data well.

Bracket delicacy (for AQI orders) Checks how rightly the model sorts AQI values into orders like Good, Moderate, or Unhealthy.

Comparative Performance Overview

Technique	Improvement Observed	Notable Tools/Models Used
Random Forest	+20-25% Accuracy	Scikit-learn, XGBoost
Neural Networks	+10-20% CodeBLEU	GitHub Copilot, keras
Decision Trees	+10-15% Interpretability	Scikit-Learn

Case Study Example

A comparative test was conducted on AQI prediction using different machine learning models.

- A fine-tuned model provided concise, optimized code with fewer logic errors.
- Structured Input: Clean, normalized data improved prediction clarity in SVM models.
- Contextual Features: Adding weather and time-based inputs helped models like Gradient Boosting and LSTM deliver more consistent results.

V. DISCUSSION AND RECOMMENDATIONS

Each machine learning technique for AQI prediction offers unique advantages, but also presents certain limitations:

- Fine-Tuning delivers high accuracy for specific regions or conditions but needs substantial training data and computing power.
- Feature Engineering (Prompt Design Equivalent) is cost-effective and improves model performance, but requires domain knowledge to select the riinputfeatures

Context Awareness (e.g., including weather or time-based factors) greatly enhances predictions, yet managing too many features can increase model complexity and risk overfitting.

Recommendations:

- Combine Fine-Tuning with Feature Selection to boost precision for localized AQI forecasting.
- Use contextual features like wind speed, humidity, and traffic data to improve real-world accuracy.
- Adopt lightweight tuning methods (e.g., grid search, random search) for limited hardware setups.
- Automate data preprocessing pipelines to ensure consistency and save time across different environments.

Conclusion

The research focused on forecasting India's Air Quality Index (AQI) for the years 2024 and 2025. A noticeable increase in AQI was seen during 2022–2023, while a sharp drop occurred in 2020 due to the nationwide Covid-19 lockdown. Afterward, AQI levels began to climb again. Among all variables, PM2.5 and PM10 played a significant role in influencing AQI, whereas meteorological parameters had limited effect. The machine learning techniques applied yielded highly accurate predictions, with Random Forest and CatBoost models achieving maximum training dataset correlations of 0.9998 and 0.9936, respectively. This indicates that ML approaches are reliable tools for AQI prediction. For broader application, these models need validation across varying environmental conditions. Moreover, testing their adaptability by implementing them in different cities or countries is crucial for evaluating their performance in diverse settings.

REFERENCES

1. MySQL 8.0 Cookbook" by Karthik Appigatla
2. Python Official Documentation – https://docs.python.org

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH

IN SCIENCE, ENGINEERING AND TECHNOLOGY

| Mobile No: +91-6381907438 | Whatsapp: +91-6381907438 | ijmrset@gmail.com |

www.ijmrset.com