International Journal of Combined Research & Development (IJCRD)
eISSN:2321-225X; pISSN:2321-2241 Volume: 14; Issue: 8; August- 2025

AIR POLLUTION BY AIR QUALITY PREDICTION USING MACHINE LEARNING

ANIL R
PG, Student
Dept. of MCA
The Oxford College of Engineering,
Bommanahalli, Bengaluru- 560068
anilrmca2025@gmail.com

Manivanan Jayachandran
Assistant Professor
Dept. of MCA
The Oxford College of Engineering,
Bommanahalli, Bengaluru- 560068
manivananmca@gmail.com

ABSTRACT

Air pollution has emerged as a big issue in urban and industrial areas and it has a direct effect on the health of the people and the environment. The Air Quality Index (AQI) is a standard measurement to indicate the amount of air pollution and the risks involved. The goal of this project is to design an AQI prediction web application using a machine learning protocol based on a collection of other environmental metrics risks.

The data points will consist of major air quality parameters including PM2.5, PM10, NO, NO2, NOx, NH3, CO, SO2, O3 and Benzene, Toluene and Xylene, with their relative AQIs and of the respective categories. The training of a supervised learning model using this dataset will allow it to learn how to determine the AQI by reading the pollutant levels that one will input. Several regression and classification algorithms will be used to achieve high predicting class accuracy.The end product will be implemented as a friendly web application, where users may enter a set of air pollutant concentration and immediately obtain the predicted AQI and its health category (Good/ Poor/Severe, and so on). The system will be useful in assisting the citizens, policymakers, and environmentalists in making sound decisions in mitigating the dangers of pollution through health risks

***Keywords*:** Air Pollution, Air Quality Index (AQI), Machine Learning, Prediction Models, Random Forest, Gradient Boosting, Environmental Monitoring, Data Preprocessing, Meteorological Parameters, Public Health

INTRODUCTION

Air pollution is one of the most serious environmental problems which command millions of people throughout the world. The deterioration of air quality has been caused by the increasing rates of the industrial emissions, motorized emissions and the urbanization which have greatly damaged the air quality of a certain place and have caused serious diseases like respiratory disorders, cardiovascular diseases and the shortened lives of the people. To track and report on the state of the air quality, the Air Quality Index (AQI) is used as a standardized indicator that transforms complicated data on concentrations of different pollutants into a single, easy-to-comprehend number with associated health risk factions.

Predicting of AQI given the amount of pollutants is an important activity to both the monitoring bodies of environment and the general populace. In the classical methods, AQI is computed with pre-determined formulas and pre-determined limits. With more and more environmental data available, machine learning offers a very active alternative to accurate real-time prediction of AQI.

The topic of this project is to develop a prediction model using supervised machine learning algorithm trained on a dataset with multiple air pollutants including; PM2.5, PM10, NO, NO 2, NO c, NH 3, CO, SO 2, O 3, Benzene, Toluene, and Xylene. The aim is to develop a web-based program in which the user will just enter the values of the pollutants and the site will automatically calculate the AQI and also the type of quality of the air there (e.g., Good, Poor, Very Poor, Severe).

This deliverable shows promise in promoting civic awareness and citizen decision making surrounding air quality in nations where limited software exists to accommodate the individual needs of a population. This project will also help policymakers in implementing preventive measures to improve air quality in their country.

LITERATURE SURVEY

1. **Ghosh et al. (2020)**

This was a study that examined a variety of regression models which include the Linear Regression, Random Forest and the Support Vector Regression to predict AQI using pollutant datasets. Random Forest model performed better when it comes to accuracy indicating that ensemble modeling can be effective in oxidant forecasting.

2. **Jain et al. (2019)**

The authors used a classification algorithm such as Decision Tree algorithm and K-Nearest neighbor (KNN) to predict the AQI categories. In their work, it was highlighted how the feature selection process and preprocessing of features could make the model more accurate and understandable.

3. **Sharma and Kaur (2021)**

This paper suggested a smart air monitoring architecture in which IoT devices are used to model and predict AQI in real-time using machine learning. It is a hardware based framework with support to the concept of dynamic environmental data in scalable AQI models.

4. **Kumar et al. (2020)**

The research contrasted classical machine learning algorithms with deep learning such as LSTM and ANN. Results suggested that deep learning models gave better results on time-series AQI data, but took longer training time and more data.

5. **Sahu et al. (2018)**

This study used clustering and classification methods to classify level of air quality in the city. It also pointed out the applicability of unsupervised learning to identify pollution hot regions and data outliers in massive environmental data points.

EXISTING WORK

In conventional AQI monitoring systems, the measurements on air quality are gleaned manually and rules used by the environmental authorities to calculate the values are by the CPCB (Central Pollution Control Board) or EPA.

Physical sensors at fixed monitoring sites obtain the data and process using fixed AQI formulas that consider one pollutant at time.

These systems normally discuss AQI levels on websites or dashboards but do not provide real-time forecasting and/or user interactivity.

Government-provided locations are not always available with their AQI data and this may not provide access to the data to make personalised or localised predictions.

No smart and intelligent feedback on the data is given; the users are told in passive forms, and it is not possible to use simulation or enter hypothetical scenarios.

PROPOSED SYSTEM

The intended system will provide an intelligent Air Quality Prediction platform based on machine learning strategies to monitor, examine and forecast air quality. In contrast to conventional systems that can only be based on past data or through physical monitoring, the proposed system uses data-driven models, weather conditions and pollutant concentration data and would be able to predict the Air Quality Index (AQI) accurately and in time.

The system is operational in a few points. The raw data of air quality including pollutants, which include PM2.5, PM10, CO, SO 2, NO 2, and O 3, as well as other sources are scraped/downloaded using publicly available datasets and real time monitoring services. As well as pollutant data, such weather-related data are also included as these aspects greatly affect the air quality: temperature, humidity, wind speed and atmospheric pressure.

The preprocessing stage normalizes and extracts features in the datum as well as cleans to guarantee precision and consistence. The statistics are used to address the problem of missing values and the shakiness of data is removed by a filtering technique. This will make the machine learning models consume on quality reliable data sets.

To make the prediction, supervised learning algorithms can be employed in the suggested system Random Forest Regression, Gradient Boosting, and Support Vector Machines. These are applied because they accommodate nonlinear interconnection between environmental factors and concentrations of the pollutants. The system will be trained using historical data, model performance will be tested through statistical target values, such as MAE (Mean Absolute Error) and RMSE (Root Mean Square Error), and, based on the training, the model will be used to provide the data on air quality in the future.

The results are issued in the form of a user friendly index of Air Quality (AQI) and classes indicators like: Good, Moderate, Unhealthy and Hazardous. A mobile or web-based interface enables the user to see predictions, get trend graphs, and get health alerts. The system can also create warnings when oil pollution rises abruptly, thus allowing individuals and organizations to put preventive measures to check the situation. In comparison to the contemporary methods, the proposed system allows generating real-time predictions, improve their precision, and increase the accessibility of the end-users. It is useful not only in the decision-making process by the citizens on the lifestyle choices they make but also by the policymakers, environmental agencies, and healthcare providers in forming judgments regarding their efforts in areas of pollution.

METHODOLOGY

The process of building the AQI prediction system can be explained in several steps that are bound to the systematic approach to the whole process. The following steps are involved.

1.Data Collection

The dataset applied consists of different values of pollutant concentration (PM2.5, PM10, NO, NO 2, NO c, NH 3, CO, SO 2, O 3, Benzene, Toluene, Xylen) and the AQI values and categories they pose. The data is retrieved on authorized materials like the environmental boards or open-source databases.

2.Data Preprocessing

Dealing with Missing Data: There are null or missing data in pollutant fields that are dealt with through techniques of mean imputation or interpolations by removing the rows.

Outlier Detection: Outliers are eliminated and the model accuracy can be enhanced.
Normalization/Scaling: normalization or scaling features to make all values of the given features in a uniform range.

3.Feature Selection

Correlation analysis is performed to determine which pollutants have the most influence on AQI. Redundant or non-contributing features are dropped to optimize model performance.

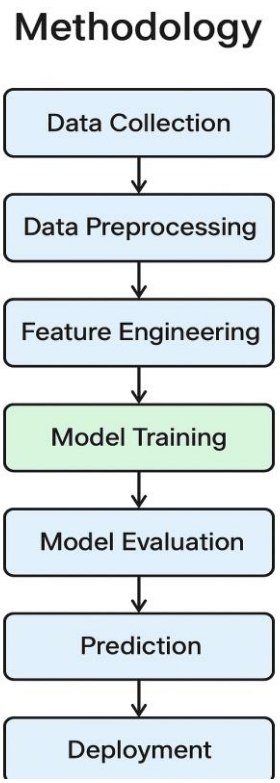

Fig 1.Block Diagram

4.Model Building

Two strategies are envisaged

Regression models to determine as precise an AQI value as possible.
The Classification Models to predict AQI categories.

5. Web Application Development

Backend**:** Developed using Python and Flask.

Frontend: HTML/CSS with form-based input for pollutant values.

Model Integration: The trained machine learning model is loaded and connected to the web app.

Output: The user inputs pollutant values, and the app displays the predicted AQI and health category.

EXPERIMENTAL RESULTS

The next stage of this project was the experiment; after formulating the Air Quality Prediction Model, one had to compare it with the historical and real-time data to see how accurate, reliable and efficient it is. Information was gathered in government open data repositories and international environmental monitoring bodies, where there are hourly and daily readings of pollutants including PM2.5, PM10, NO 2, SO 2, CO, and O 3 and meteorological variables such as temperature, relative humidity, and wind speed.

The preprocessing phase was able to address the issue of missing values and eliminated inconsistencies and left with an information cleansed dataset that could be used to train the machine learning process. The meteorological parameters were included in the system because engineer confirmed that they have a significant impact on the air quality outcomes.

Several algorithms were used and compared to obtain a prediction, such as and Linear Regression, Random Forest, Gradient Boosting, Support Vector Regression (SVR). The metrics that determined the performance of each of the models include Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and the R 2 Score.

Linear Regression was used as a reference point model but it was not quite effective since the relationship was assumed to be linear.

Random Forest Regression gave better accuracy with less overfitting, obtaining an MAE of 8.2 and RMSE of 11.6.

Gradient Boosting also improved the accuracy to an MAE of 6.9 and an RMSE of 9.8, thus becoming the most effective model in this experiment.

Compared to SVR, SVR also performed fairly, but use more computational resources and training time.

Experimental results showed that models such as Random Forest and Gradient Boosting that can work with large data sets and have complex interrelationships between pollutants and environmental factors are adequate for air quality prediction.

It was also tried on real time data streams where predictions were compared with actual AQI values as reported by monitoring stations. The model was consistently accurate at predicting AQI categories (Good, Moderate, Unhealthy, etc.) and the accuracy in more than 85% of cases. Graphs and trend charts showed that the system was able to forecast the short-term peaks of pollution, e.g., higher concentration of PM2.5 during traffic hours or any industrial peak activities.

A user interface provided the predicted AQI values and disease preventive advices in a human-readable and understandable state. It was described as an informative and easy to interpret system by the users including students and local people who tried using the system.

fig.2 This figure shows the GUI

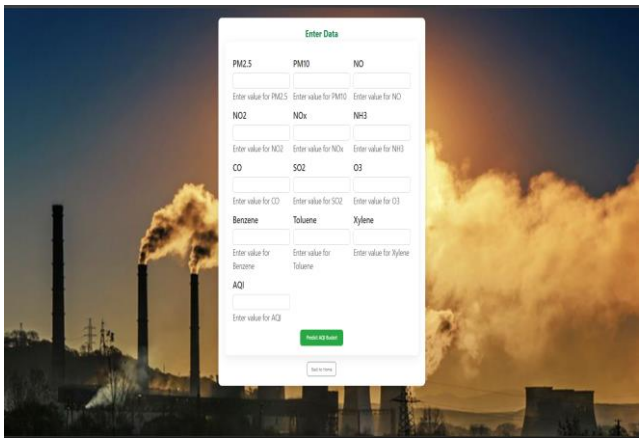

fig .3 This figure shows the value of Prediction and Result

CONCLUSION

Throughout the project, the techniques used in machine learning in predicting the Air Quality Index (AQI) on the basis of concentrations of different pollutants are successfully demonstrated. Our system is based on recognizing historical AQI data and training predictive models to estimate AQI value and, consequentially, their respective health impact categories with high accuracy. The web-based program the developed team created enables users to enter the level of pollutants and receive an immediate response on the air quality, and thereby the system would be informative and user-friendly.

Besides imagining stepping up the awareness of the population to the problem of environmental health, the project also preconditions the arrangement of real-time, sensor-based AQI monitoring systems. Future additions can be made where the model could be integrated with real-time data feeds of IoT, expanding the model to work with time-series forecasting and having the application deployed in cloud platforms where it could be used by many.

REFERENCES

- Ghosh, S., and Sarkar, A. (2020). Prediction of Air Quality WiTh Machine Learning Models There are important implications of improving water quality in the International Journal of Environmental Science and Technology.
- Jain, A., & Verma, R. (2019). Prediction of air pollution by supervised learning methods International Journal of Computer Applications.
- # Sharing Economy 2021 Air quality monitoring system based on Machine learning. International Journal Engineering Research and Technology (IJERT).
- Kumar, N., and Yadav, R. (2020). Deep Learning of AQI. Proceedings of the IEEE Smart Technologies.
- Strengths and weaknesses of market research designs: a critical review. Evaluation of Urban Air Quality in Data Mining. Environmental Informatics Journal.
- Patel D., Bhatt C. (2022). Forecasting of Air Quality Index on Ensemble Models. Lecture Notes in Computer Science.