

# AIR QUALITY PREDICTION USING DIFFERENT MACHINE LEARNING MODELS

Aditya Singh  
Computer Science and Engineering  
Graphic Era Hill University,  
Dehradun, India  
[adityasingh776433@gmail.com](mailto:adityasingh776433@gmail.com)

Satvik Vats, SMIEEE  
Computer Science and Engineering,  
Graphic Era Hill University;  
Adjunct professor,  
Graphic Era Deemed to be University  
Dehradun, India.  
[svats@gehu.ac.in](mailto:svats@gehu.ac.in)

Lakshit Sharma  
Computer Science and Engineering  
Graphic Era Hill University,  
Dehradun, India  
[lakshitsharma1852004@gmail.com](mailto:lakshitsharma1852004@gmail.com)

Vikrant Sharma, SMIEEE  
Computer Science and Engineering,  
Graphic Era Hill University;  
Adjunct professor,  
Graphic Era Deemed to be University,  
Dehradun, India.  
[vsharma@gehu.ac.in](mailto:vsharma@gehu.ac.in)

**Abstract**— The cost of air pollution in Indian cities has increased significantly in both rural and urban areas. Air pollution can affect body health. AQI is a term used to understand the air quality in a city or country.

AQI stands for Air Quality Index. AQI is used to provide an easy-to-understand way to measure daily weather and its effects on health. AQI usually ranges from 0 to 500; higher numbers indicate worse air quality. The index is divided into several categories, each corresponding to different health problems.

1. 0-50: Good
2. 51-100: Moderate
3. 101-200: Unhealthy
4. 201-300: Very Unhealthy
5. 301-500: Hazardous.

The aim of this paper is to find the most effective machine learning model of forecasting AQI to help climate monitoring. The most efficient method can be refined to find the optimal method solution. Therefore, the work in this paper involves intensive research and applying machine learning model that is the best solution for air quality problems.

In Air Quality Prediction we have used Logistic Regression, Decision Tree Classifier, Random Forest Classifier, K-Nearest Neighbor Models.

It is noted that Random Forest Classifier has highest value of Accuracy Test and Logistic Regression with the lowest value of Accuracy Test. Random Forest Classifier has the highest value of Kappa Score and Logistic Regression has lowest value of Kappa Score. Random Forest Regressor has best R Squared Test Score and Linear Regression has poorest R Squared Test Score.

**Keywords:** Machine Learning, AQI, Model, Logistic Regression, Decision Tree, Random Forest, K-Neighbor.

number of AQI indicates highly contaminated air which negatively impacts our health and fitness. Data miners captured AQI of different cities of our countries and transformed it into a dataset which will help us to predict where quality of air is good or bad. Dataset of Air Quality of India contain different parameters SO<sub>2</sub>, NO<sub>2</sub>, RSPM, SPM to calculate AQI of a particular city or of area. Different regression models will be applied to the dataset to know which provides best accuracy and then predict the AQI by picking values from the dataset or entering values ourself suffer from respiratory diseases and death. According to scientific evidence, air pollution poses the greatest environmental risk. As a result of toxic gas emissions caused by rapid industrialization, the population has increased dramatically. Our health suffers greatly as a result of air pollution with hazardous substances. Due to this uncontrolled pollution, the air quality has decreased significantly.

AQI is a numerical index used to measure and express the level of air pollution. There is total 18 parameters to calculate AQI of Air. Which are NO<sub>2</sub>(nitrogen dioxide), SO<sub>2</sub>(sulfur dioxide), CO (carbon monoxide), O<sub>3</sub>(ozone), PM<sub>10</sub>(particulate matter having diameter ≤ 10 microns), PM<sub>2.5</sub>(particulate matter having diameter ≤ 2.5 microns), NH<sub>3</sub> (ammonia), benzene, PM<sub>10</sub>, PM<sub>2.5</sub>, SO<sub>2</sub>, NO<sub>2</sub>, CO, and O<sub>3</sub> are used to calculate the air quality index (AQI).

A greater number of AQI indicates highly contaminated air which negatively impacts our health and fitness. Data miners captured AQI of different cities of our countries and transformed it into a dataset which will help us to predict where quality of air is good or bad. Dataset of Air Quality of India contain different parameters SO<sub>2</sub>, NO<sub>2</sub>, RSPM, SPM to calculate AQI of a particular city or of area. Different regression models will be applied to the dataset to know which provides best accuracy and then predict the AQI by picking values from the dataset or entering values ourself.

## I. INTRODUCTION

Due to polluted air, millions of people around the world

## II. LITERATURE SURVEY

A study was done in 2019, they first examined the relationship between various weather factors such as AQI, PM2.5 concentration and total NOx (nitrogen oxide) concentration.[1]

In 2020, they developed a prediction model using random forest regression (RFR) and support vector regression (SVR), and finally evaluated the performance of the regression model using RMSE, coefficient of determination (R-Square), and correlation coefficient R. Measure air pollution and common problems and predict air pollution [2].

In 2021, According to the study, hourly air pollutants in California such as carbon monoxide, sulfur dioxide, nitrogen dioxide, ground-level ozone and chemical 2.5 can be estimated using SVR along with hourly AQI. EPA provides an unprecedented data set with 94.1% accuracy for six AQI categories. MLR and monitoring systems were used to estimate AQI. Use a variety of metrics to measure performance. Second, the ARIMA time series model is used to predict future AQI. Both models have proven to be accurate and useful in predicting AQI [3].

A study was done in 2015, A widely used model using non- linear logistic regression and kriging to predict air pollution in Mumbai and Navi Mumbai. A high R value indicates reasonable agreement between expectations and predictions. ANN outperforms simple regression models in terms of R value and prediction [4].

In 2020, they predicted AQI concentration based on PM2.5, PM10, SO2, NO2 and other parameters. Briefly, among the linear regression, decision tree regression, SVR and RFR algorithms, the random forest regression algorithm with the best accuracy in the evaluation data is 0.99985, the average power method of the least squares method is 0.00013 and the average absolute error 0.003730 [5].

In 2019, Using previous year's data to predict AQI and using gradient descent to predict next year will compound the problem. They improve the performance of the model by applying the predictive value to the prediction problem, thus outperforming the regression model. They also used the AHP MCDM technique to evaluate the requirements based on the similarity of the best options [6].

A study was done in 2018, Logistic regression [7] was used to determine whether air/environmental samples taken from a particular city are polluted or not. The system tests PM2.5 levels and detects air quality based on previous PM2.5 readings. The results show that logistic regression and autoregression can be effectively used to diagnose future climate and predict PM2.5 levels.

In 2017, This study [8] used 6 years of weather and climate data to provide machine learning to predict PM2.5

concentrations based on wind (speed and direction) and precipitation. The results of the classification model showed good performance in classifying low (10 g/m<sup>3</sup>) and high (>25 g/m<sup>3</sup>) PM2.5 concentrations, as well as low (10 g/m<sup>3</sup>) and medium PM2.5 concentrations.

In 2019, A widely used model that uses neural networks and kriging to predict pollution levels in Mumbai and Navi Mumbai based on historical data from the Ministry of Health and Safety [9]. The proposed model was implemented and tested using MATLAB application for ANN and R application for kriging. The system helps analyze pollution data and predict future pollution.

In 2019, Time series analysis is also used to analyze future data regarding extreme weather conditions. An effective strategy for hourly weather forecasting using LSTM based deep RNN was investigated to predict AQI in Delhi. The results are accurate regardless of the hourly forecast. As shown in research [10].

## III. DATASET SELECTION & SAMPLE DATA

We have selected the dataset from Kaggle:

<https://www.kaggle.com/datasets/shrutibhargava94/india-air-quality-data/data>

The dataset contains Stncode, Sampling Date, State, Location, Agency, Type, SO2, NO2, RSPM, SPM, Location Monitoring Station, PM 2.5, Date.

The dataset has hourly and daily air quality and AQI (air quality index) data from numerous stations in several Indian cities. The data are for the years 1990 through 2015. The original dataset included 435741 rows and 13 columns, which included all of the cities listed below. The cities are given below Hyderabad, Pantancheru, Tirupati, Visakhapatnam, Ramagundam, Vijayawada, Kurnool, etc.

The dataset has 6 columns of Quantitative Data which are Stncode, SO2, NO2, RSPM, SPM PM 2.5 and 1 column with Date Data which is date. We dropped 5 columns from dataset which are Agency, Stncode, Date, Sampling Date, Location Monitoring Station because there are many missing values and they are not much useful in finding AQI. After Cleaning the dataset, we get 8 columns having int 64 type of data.

Where no. of rows from each state are as: Maharashtra: 60384, Uttar Pradesh: 42816, Andhra Pradesh: 26368, Punjab: 25634, Rajasthan: 25589, Kerala: 24728, Himachal Pradesh: 22896, West Benga: 22463, Gujarat: 21279, Tamil Nadu: 20597, Madhya Pradesh: 19920, Assam: 19361, Odisha: 19279, Karnataka: 17119, Delhi: 8551, Chandigarh: 8520, Chhattisgarh: 7831, Goa: 6206, Jharkhand: 5968, Mizoram: 5338, Telangana: 3978, Meghalaya: 3853, Puducherry: 3785, Haryana: 3420, Nagaland: 2463, Sikkim: 1, Andaman-and-Nicobar Islands: 1, Lakshadweep: 1, Tripura: 1.

No. of rows from each type: Residential, Rural and other Areas: 184407, Industrial Area: 96091, Residential and others: 86791,

Industrial Area: 51747, Sensitive Area: 8980, Sensitive Areas: 5536, RIRUO: 1304, Sensitive: 495, Industrial: 233, Residential: 158.

#### IV. METHODOLOGY

We have opted for several machine learning models- logistic regression, decision tree classifier, random forest classifier, K-Neighbor classifier to find AQI.

Logistic Regression Model advantages:

1. Easy to understand and use.
2. Good binary distribution.
3. Considering the result.
4. Suitable for small files.
5. Less chance of overfitting.

Decision Tree Model advantages:

1. Interpretability
2. Visualization
3. Working with numerical and categorical data.
4. Addresses non-linearity.
5. Auto variable selection option.

Random Forest Model advantages:

1. Higher Accuracy
2. Very less chances of overfitting.
3. Handles large dataset very easily.
4. Easily handles high-dimensional data.
5. Can be adapted by different functions.

Random Forest Model advantages:

1. Easier Implementation.
2. It is a non-parametric method.
3. No training phase required.
4. Easily adapt to new data.

We have used the following libraries:

NumPy, pandas, seaborn, matplotlib, sklearn.

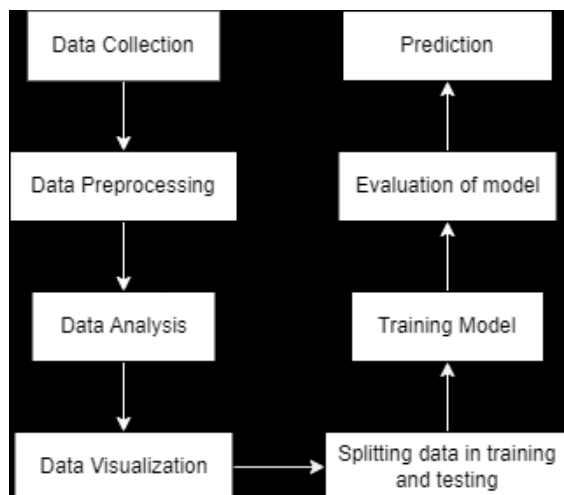


Figure 1: Methodology followed

#### A: Data Collection

By analysing figure 1 we can see that our first step is to collect data from the dataset, so for that we have used Kaggle website for our air quality dataset then load the csv file in data frame.

#### B: Data Preprocessing

In this step we will do statistical analysis and cleaning of the dataset where we will count the no. of rows in the dataset, check whether there is a null observation or not, minimum and maximum value of each column, mean value of the column, standard deviation, then calculate the value of 25%, 50%, 75% of data of a column. After that we cleaned our dataset by deleting unnecessary columns which contains large number of null values.

Then searched the amount of not unique value in the dataset so that we can get the information about the no. of rows that are duplicate or gives same information.

#### C: Data Analysis

By seeing figure 1 we can say that after data preprocessing, we have to do data analysis on the dataset. In this step we will fetch the information about the dataset. First plotted the histogram of no. of rows or information each state contains, then plotted histogram of no. of types of area each observation has. After that calculated the amount of SO<sub>2</sub>, NO<sub>2</sub>, RSPM, SPM, PM 2.5.

Then finally sorted the amount of SO<sub>2</sub> and NO<sub>2</sub> each state contains

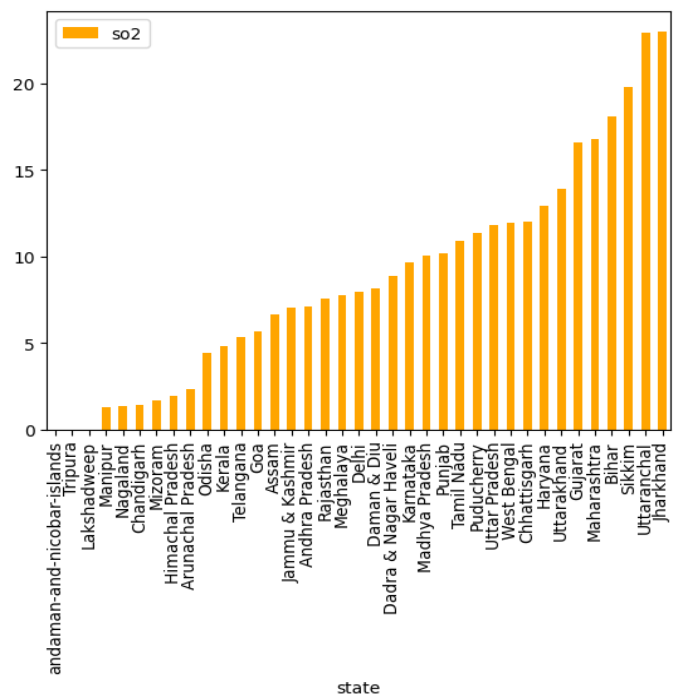


Figure 2: Amount of SO<sub>2</sub> in each state

Then searched the amount of NO<sub>2</sub> in air in each state particularly,

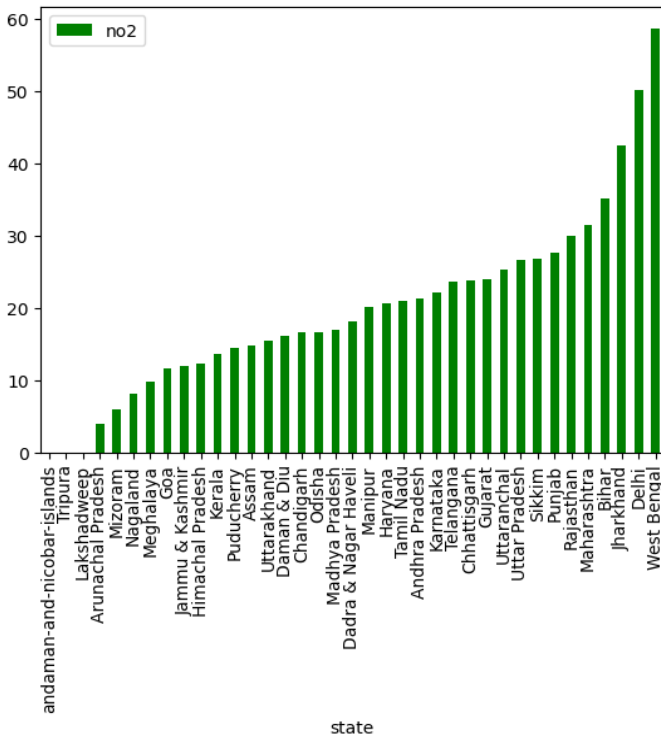


Figure 3: Amount of SO2 in each state

## V. RESULT

STATE	LOCATION	SO2	NO2	RSPM	SPM	AQI
Andhra Pradesh	Hyderabad	8.1	30.2	91	206	Poor
Assam	Bongaigoan	2.7	7.3	21	36	Good
Delhi	Delhi	4.2	54.8	221	460	Hazardous
Gujarat	Ahemdabad	10	13	73	24	Good

Figure 4: AQI using Logistic Regression

STATE	LOCATION	SO2	NO2	RSPM	SPM	AQI
Andhra Pradesh	Hyderabad	8.1	30.2	91	206	Unhealthy
Assam	Bongaigoan	2.7	7.3	21	36	Good
Delhi	Delhi	4.2	54.8	221	460	Hazardous
Gujarat	Ahemdabad	10	13	73	24	Good

Figure 5: AQI using Decision Tree Classifier

STATE	LOCATION	SO2	NO2	RSPM	SPM	AQI
Andhra Pradesh	Hyderabad	8.1	30.2	91	206	Unhealthy
Assam	Bongaigoan	2.7	7.3	21	36	Good
Delhi	Delhi	4.2	54.8	221	460	Hazardous
Gujarat	Ahemdabad	10	13	73	24	Good

Figure 6: AQI using Random Forest Classifier

STATE	LOCATION	SO2	NO2	RSPM	SPM	AQI
Andhra Pradesh	Hyderabad	8.1	30.2	91	206	Unhealthy
Assam	Bongaigoan	2.7	7.3	21	36	Good
Delhi	Delhi	4.2	54.8	221	460	Hazardous
Gujarat	Ahemdabad	10	13	73	24	Good

Figure 7: AQI using K-Nearest Neighbor

### D: Data visualization

In this step we use libraries like seaborn, pyplot to plot the graphs and charts to know the information or facts and figure through visualization because it is easier to visualize through diagrams (charts and figures).

### E: Splitting data in training and testing

After data visualization we will divide our dataset into training and testing part as we have to train out model. So, we have divided our dataset into 80% training=0.8 and 20% in testing=0.2. We have taken 4 variables X\_train, X\_test, Y\_train and Y\_test and then inserted all data in X\_train and all testing data into Y\_train and all target into Y\_train and testing target into Y\_test.

### F: Training Model

In this step we have trained our four model logistic regression, decision tree, random forest and k-nearest neighbor. Loaded all of the dataset with X\_train, X\_test, Y\_test, Y\_train for training of the model.

### G: Evaluation of Model

In this step we evaluated the model that we have trained on some of the parameters which are- accuracy of training data, accuracy of testing data and kappa score through which we can get to know which model performed best.

### H: Prediction

After all of the above steps we can predict the quality of the air by giving observation to the model.

ALGORITHM NAME	ACCURACY ON TEST DATA
Logistic Regression	0.72712
Decision Tree Classifier	0.99980
Random Forest Classifier	0.99981
K-Nearest Neighbor	0.99671

Figure 8: Name of Algorithm with Accuracy on Testing Data

As you can see Decision Tree Classifier and Random Forest Classifier gives best accuracy to find AQI.

## VI. CONCLUSION

While there are several ways to find air quality of the atmosphere but there are several challenges that can occur while finding AQI (Air Quality Index)-

1. **Data quality:** The accuracy of air quality depends on the quality and quantity of input data. Obtaining detailed and reliable data from various sources such as weather stations, air quality monitoring stations, satellite observations and ground sensors can be difficult due to inconsistencies, information gaps and errors.
2. **Spatial and temporal variation:** Air quality may vary from place to place depending on factors such as local emissions, atmospheric conditions and geography. Forecast models should include this variable to provide accurate and localized predictions.
3. **Complexity of atmospheric processes:** The atmosphere is controlled by physical and chemical processes, including weather conditions, pollutant emissions, chemical reactions, and transportation. Effective modeling of these processes requires complex mathematical models and computational tools.
4. **Estimating emissions:** Estimating emissions from many sources, including industry, transportation, agriculture, and fires, is difficult due to incomplete data, different emission factors, and uncertainty at the operation level. The right emissions products are essential for improving air quality.
5. **Model validation and evaluation:** Validation and evaluation of air quality models should compare model predictions with observations from monitoring stations. However, obtaining good clinical data for practical purposes can be difficult, especially in underserved areas.

6. **Multiple data integration models:** Good weather forecasts often require integration of data from different sources, including meteorological data, satellite observations, land use data and exports. Integrating these different data sets poses a major challenge while accounting for differences in spatial and temporal resolution.
7. **Model complexity and interpretability:** Assessing model complexity and interpretability is important for the development of good climate prediction models. Although complex models can accurately capture atmospheric processes, they can be difficult and expensive to interpret.

Solving these challenges will significantly increase the accuracy of finding AQI of the air using different machine learning models.

## VII. REFERENCES

1. H. Liu, Q. Li, D. Yu, and Y. Gu, "Air quality index and air pollutant concentration prediction based on machine learning algorithms," *Applied Sciences*, vol. 9, p. 4069, 2019.  
View at: [Publisher Site](#) | [Google Scholar](#)
2. M. Castelli, F. M. Clemente, A. Popovic, S. Silva, and L. Vanneschi, "A machine learning approach to predict air quality in California," *Complexity*, vol. 2020, Article ID 8049504, 23 pages, 2020.  
View at: [Publisher Site](#) | [Google Scholar](#)
3. G. Mani, J. K. Viswanadhapalli, and A. A. Stonie, "Prediction and forecasting of air quality index in Chennai using regression and ARIMA time series models," *Journal of Engineering Research*, vol. 9, 2021.  
View at: [Publisher Site](#) | [Google Scholar](#)
4. S. V. Kottur and S. S. Mantha, "An integrated model using Artificial Neural Network (ANN) and Kriging for forecasting air pollutants using meteorological data," *Int. J. Adv. Res. Comput. Commun. Eng.*, vol. 4, pp. 146–152, 2015.  
View at: [Publisher Site](#) | [Google Scholar](#)
5. S. Halsana, "Air quality prediction model using supervised machine learning algorithms," *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, vol. 8, pp. 190–201, 2020.  
View at: [Publisher Site](#) | [Google Scholar](#)

6. A. G. Soundari, J. Gnana, and A. C. Akshaya, "Indian air quality prediction and analysis using machine learning," *International Journal of Applied Engineering Research*, vol. 14, p. 11, 2019.

View at: [Google Scholar](#)

7. C. R. Aditya, C. R. Deshmukh, N. D K, P. Gandhi, and V. astu, "Detection and prediction of air pollution using machine learning models," *International Journal of Engineering Trends and Technology*, vol. 59, no. 4, pp. 204–207, 2018.

View at: [Publisher Site](#) | [Google Scholar](#)

8. J. Kleine Deters, R. Zalakeviciute, M. Gonzalez, and Y. Rybarczyk, "Modeling PM2. 5 urban pollution using machine learning and selected meteorological parameters," *Journal of Electrical and Computer Engineering*, vol. 2017, Article ID 5106045, 14 pages, 2017.

View at: [Publisher Site](#) | [Google Scholar](#)

9. P. Bhalgat, S. Pitale, and S. Bhoite, "Air quality prediction using machine learning algorithms," *International Journal of Computer Applications Technology and Research*, vol. 8, pp. 367–370, 2019.

View at: [Publisher Site](#) | [Google Scholar](#)

10. M. Bansal, "Air quality index prediction of Delhi using LSTM," *Int. J. Emerg. Trends Technol. Comput. Sci*, vol. 8, pp. 59–68, 2019.

View at: [Google Scholar](#)

