PREDICTION AND ANALYSIS MODEL FOR THE AIR QUALITY INDEX OF BENGALURU CITY USING REGRESSION ALGORITHMS

Jayadeva Javali[*], Sharanya K[**], Santhosh Rebello[***]

Affiliation:

1. Post Graduate Department of Information Technology, St. Aloysius College (AIMIT) (Autonomous), Mangalore, Karnataka 575022.
2. Department of Big Data Analytics, St. Aloysius College (AIMIT) (Autonomous), Mangalore, Karnataka 575022.
3. Department of Big Data Analytics, St. Aloysius College (AIMIT) (Autonomous), Mangalore, Karnataka 575022.

Abstract:

Air is a mixture of many gasses and dust particles. It is a gas that living organisms breathe to survive. Presently, there are many types of researches that claim the poor quality of air around us. Air Quality Index (AQI) is the measure used by the government to communicate to the public how polluted the air is and how it is forecasted to be in the future. High AQI increases the risk of public health. The computation of the air quality index requires the concentration of various gases like PM2.5, PM10, NO, O3 etc.

In this research paper, a basic analysis of the data is done to understand the data and Machine Learning Techniques like regression algorithms have been applied to predict the future AQI. Our model will be capable of successfully predicting the per day air quality index of Bengaluru city given the daily pollutant levels in the air. With the diminishing quality of the AQI, this paper tries to analyze, predict and provide possible solutions for the betterment of the future.

Keywords:

Air Quality Index, contaminated, pollutant, monitoring stations, public health, ranking, raw data, preprocessing, data validation, machine learning, outlier, regression, RMSE, R2, MAE, accuracy, relationship, 3 R's, EVs.

1. Introduction

Air Quality Index (AQI) is a metric that government agencies use to inform the public about how contaminated the air is now or will become in the future. As the AQI grows, so do the threats to public health. Air quality indices differ by country and correspond to distinct national air quality standards. The concentration of an air pollutant during a defined averaging period, collected from an air monitor or model, is required for AQI computation. When concentration and time are added together, the total dosage of the air pollutant is calculated.

The AQI results are divided into predetermined ranges, each with its color code, description, and standardized health recommendation.

The deterioration of air quality is merely one of the negative consequences of pollutants emitted into the atmosphere. Over the last few decades, other negative repercussions such as acid rain, global warming, aerosol production, and photochemical smog have also increased.

Ground-level ozone (O3), Sulphur dioxide (SO2), particulates matter (PM10 and PM2.5), carbon monoxide (CO), carbon dioxide (CO2), and nitrogen dioxide (NO2) are all monitored by the Environmental Protection Agency (EPA) (NO2). These compounds are part of the Air Quality Index (AQI), a popular metric that indicates how clean or polluted the air is now or will be in a certain area. As the AQI rises, a larger proportion of the population becomes exposed. Air quality indices vary by country and correspond to distinct air quality standards.

India, as the world's fastest-growing industrial nation, is emitting unprecedented levels of pollutants, including CO2, PM2.5, and other dangerous airborne contaminants. The impact of pollutants on specific regions is measured by the air quality of a state or country. According to the Indian air quality standard, contaminants are indexed in terms of their scale, and these air quality indices reflect the amounts of main pollutants in the atmosphere. There are a variety of

atmospheric gases that pollute our environment. At various degrees, each pollution has its indicator and scale. The key pollutants such as (no2, so2, rspm, spm) indices AQI are obtained, and the data can be categorized based on the limits using this individual AQI.
Every accessible data point in the dataset has been used to create a model that predicts the air quality index. We can identify the principal pollution-causing pollutant and the location severely affected by the pollutant across India by predicting the air quality index. Air is everywhere, and its impact is as well. Air pollution has a negative impact on human health, buildings, monuments, plants, ecosystems, and so on. Particulate matter absorbs or reflects sunlight, affecting cloud formation and rainfall patterns. Polluted air has been related to climate.

2. Problem Statement

Most authorized centers have failed to record real-time data on air pollution due to a lack of monitoring and maintenance.

AQI Category	AQI Range
Good	0-50
Satisfactory	51-100
Moderate	101-200
Poor	201-300
Very Poor	301-400
Severe	401-500

Fig 1: AQI Range list.

The National Air Quality Index (NAQI) is a tool that uses numbers to simplify air quality data by categorizing pollution levels into six categories—good, satisfactory, moderate, poor, very poor, and severe—and assigning a color code based on how damaging pollution in a given area is. Each pollutant—PM2.5, PM10, NO2, CO, and Ozone—is given an air quality index (AQI), and then a daily total AQI is given, showing the worst pollutant value for that area. This allows users to see how awful the pollution is in their neighborhood and which sites they should avoid on that particular day. With the availability of a monitoring tool like this, there is no doubt that awareness will rise.

But what is missing is that no one knows the next step. The administration has made no mention of a bold action plan that would advise citizens on what precautions to take based on pollution levels and how to address the larger issue of air pollution.

The monitoring stations in the selected locations must be operational for this AQI to function effectively. While many people had high hopes for this AQI, it has already begun to display flaws as a result of poor upkeep.

Statement: There is no doubt that the air pollution has a negative influence on the environment and climate. A rating of 201 to 300 on the Air Quality Index (AQI) indicates very dangerous air quality for survival, with high levels of health risk.
As the number of automobiles in our city grows, so does air pollution, which has a negative impact on the environment. Our mission is to predict the Air Quality Index using historical data.

2.1 Project Objectives

The project's goals are to:
(i) Inform the public about the overall state of air quality using an easy-to-understand summation parameter;
(ii) Inform residents about the health effects of air pollution exposure; and
(iii) Rank cities/towns for prioritizing measures using an AQI measure.

Air quality standards are the most important legal foundation for air pollution control. An air quality standard is a statement of a regulatory authority's enforceable level of air quality. The purpose of standard development is to provide a reason for protecting public health from the detrimental effects of air pollutants, eliminating or lowering hazardous air pollution exposure, and guiding national/local authorities in pollution control decisions. CPCB issued

a new set of Indian National Air Quality Standards (INAQS) for 12 criteria with these goals in mind (air constituents). Except for CO and O3, the first eight parameters have both short-term (1/8/24 hours) and annual standards, whereas the other four parameters have just annual standards.

2.2 Study Area

The dataset comprises air quality data and AQI (Air Quality Index) for numerous stations across India at the hourly and daily levels. The data is available on the Central Pollution Control Board's website, https://cpcb.nic.in/, which is the government of India's official portal. They also offer an app that monitors air quality in real-time: https://app.cpcbccr.com/AQI India/

Bengaluru, India's cultural, educational, industrial, and administrative capital, is situated between 12º 59 ' north and 77º 35' east longitudes, at an elevation of 900 meters above sea level. The average rainfall is 1286.6mm, with temperatures ranging from 7.8°C to 38.9°C. In the last few decades, its urban area has risen by 466 per cent, with a decadal population growth of 51.39 per cent. To accommodate such a huge population, the city borders have been increased at the expense of adjacent villages10 and the illegal conversion of green belt land11. The purpose of this paper is to look at the changes in air quality over the last 10 years and the variables that have contributed to those changes. During this time, Bengaluru's image shifted from that of a garden city to that of a major IT hub.

Ambient Air Quality Monitoring Station 1	BTM Layout, Bengaluru - CPCB
Ambient Air Quality Monitoring Station 2	BWSSB Kadabesanahalli, Bengaluru - CPCB
Ambient Air Quality Monitoring Station 3	Bapuji Nagar, Bengaluru - KSPCB
Ambient Air Quality Monitoring Station 4	City Railway Station, Bengaluru - KSPCB
Ambient Air Quality Monitoring Station 5	Hebbal, Bengaluru - KSPCB
Ambient Air Quality Monitoring Station 6	Hombegowda Nagar, Bengaluru - KSPCB
Ambient Air Quality Monitoring Station 7	Jayanagar 5th Block, Bengaluru - KSPCB
Ambient Air Quality Monitoring Station 8	Peenya, Bengaluru - CPCB
Ambient Air Quality Monitoring Station 9	Sanegurava Halli, Bengaluru - KSPCB
Ambient Air Quality Monitoring Station 10	Silk Board, Bengaluru - KSPCB

Fig 2: AQI Monitoring Stations, Bengaluru City.

2.3 Bengaluru Air Pollution Measurement

Regularly, air pollutants are measured under the National Ambient Air Quality Programme (NAMP) by monitoring stations and the KSPCB (Karnataka State Pollution Control Board) publishes an annual average of pollutants such as Sulphur dioxide (SO2), nitrogen dioxide (NO2), and particulate matter with aerodynamic diameters less than ten millimeters (PM10 or RSPM) on its website. These data were used to determine the Exceedance Factor for six stations. These stations were chosen since measurements for them are accessible for the longest time (2015-20), and they also represent mixed urban [Yeshwanthpur (YPR) and AMCO Batteries (AMCO)], industrial [Peenya, KHB, and Graphite India (GI)], and sensitive area [Victoria Hospital (VH)] environments.

AQI Category (Range)	PM_{10} 24-hr	$PM_{2.5}$ 24-hr	NO_2 24-hr	O_3 8-hr	CO 8-hr (mg/m^3)	SO_2 24-hr	NH_3 24-hr	Pb 24-hr
Good (0-50)	0-50	0-30	0-40	0-50	0-1.0	0-40	0-200	0-0.5
Satisfactory (51-100)	51-100	31-60	41-80	51-100	1.1-2.0	41-80	201-400	0.6 –1.0
Moderate (101-200)	101-250	61-90	81-180	101-168	2.1- 10	81-380	401-800	1.1-2.0
Poor (201-300)	251-350	91-120	181-280	169-208	10.1-17	381-800	801-1200	2.1-3.0
Very poor (301-400)	351-430	121-250	281-400	209-748*	17.1-34	801-1600	1201-1800	3.1-3.5
Severe (401-500)	430 +	250+	400+	748+*	34+	1600+	1800+	3.5+

Fig 3: AQI Components Range.

AQI	Associated Health Impacts
Good (0–50)	Minimal Impact
Satisfactory (51–100)	May cause minor breathing discomfort to sensitive people
Moderate (101–200)	May cause breathing discomfort to the people with lung disease such as asthma and discomfort to people with heart disease, children and older adults
Poor (201–300)	May cause breathing discomfort to people on prolonged exposure and discomfort to people with heart disease with short exposure
Very Poor (301–400)	May cause respiratory illness to the people on prolonged exposure. Effect may be more pronounced in people with lung and heart diseases
Severe (401-500)	May cause respiratory effects even on healthy people and serious health impacts on people with lung/heart diseases. The health impacts may be experienced even during light physical activity

Fig 4: Associated AQI health Impacts.

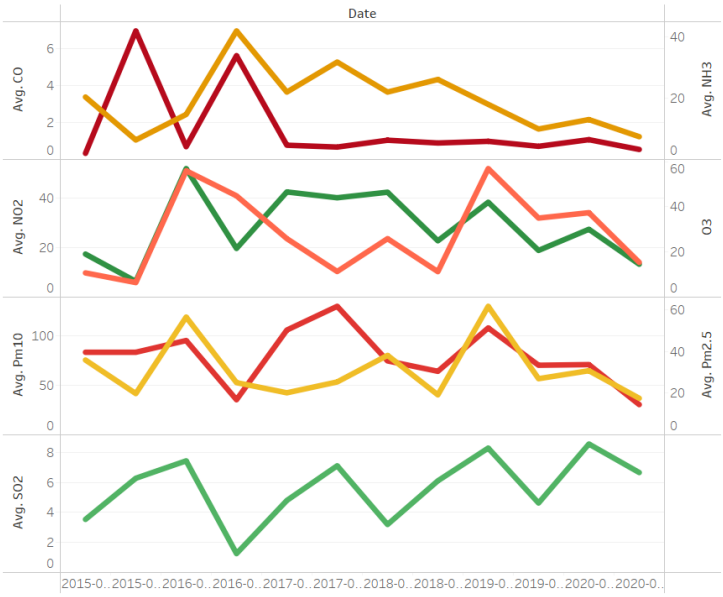

Fig 5: Average composition of gases in air measured over every six months from 2015-2020.

2.4 AQI based ranking of cities

The following is the outline of the procedures that can be used for ranking the cities:

- Collect air quality data for all eight or a minimum of three pollutants for averaging time periods as specified in INAQS (Indian National Air Quality Standards).
- AQI based Ranking should be carried out for cities with a population of 1 million or larger.
- The ranking should be done every six months on January 1st and July 1st.
- Determine the area and corresponding population of the domain (2km x 2km) in which monitoring is being carried out in the city.
- For ranking purposes, both online monitoring and manual monitoring sites should be accounted. AQI data from a minimum of three stations should be available in the city.

3 Methodologies:

3.1 Data Sources

Secondary data

- cpcb.nic.in
  The Central Pollution Control Board of India is statutory organization under Ministry of Environment, Forest and Climate Change. CPCB runs nationwide programs of ambient air quality monitoring known as National Air Quality Monitoring Programme (NAMP). The site provides real time as well as historical data for usage of public purposes.
- Kaggle
  It is an Online Community of Data Scientists and Machine Learning Practitioners. It helps user in finding and publishing datasets. It also helps user in utilizing a web-based data-science environment to collaborate, explore and build models to solve challenges.

3.2 Data Preprocessing:

Data Preprocessing is a data mining technique for transforming raw data into a usable and efficient format. Raw data is frequently incomplete and formatted inconsistently. The success of every project involving data analytics is directly proportional to the quality of data preparation. Data validation and data imputation are both parts of the preprocessing process. 'The purpose of data validation is to determine whether the data is comprehensive and accurate. Both database-driven and rules-based applications utilize data preprocessing. Data preprocessing is crucial in machine learning (ML) processes to ensure that big datasets are prepared in such a way that the data they contain can be processed and analyzed by learning algorithms.

3.3 Dealing with Missing Values:

The practice of detecting and correcting inaccurate/incorrect data in a dataset is known as data cleaning. One of the processes that are required is to address the missing values in the dataset. Numerous datasets will include many missing values in real life, therefore dealing with them is a crucial step.

Why is it necessary to fill in the blanks? Because passing NaN numbers into most of the machine learning models you want to utilize would result in an error. The simplest solution is to simply fill them with 0, however, this will greatly lower your model's

accuracy. There are numerous methods for filling in missing values. In this project, we have imputed the missing values using the Mean imputation method as the data is continuous in nature. This approach can be applied with features that are independent of each other. Though this can be extended only in a numerical dataset, it depends on the nature of that particular feature. All the missing values are replaced with the mean of the respective features

3.4 Outlier Treatment:

An outlier is a data point that is significantly different from the rest of the population's values. We frequently wish to make assumptions about a particular group of people. Extreme values, on the other hand, might have a major impact on data or machine learning model findings. Abnormal observations can be seen as part of different populations with outlier detection and treatment, ensuring stable findings for the population of interest.

Outliers, once detected, can offer surprising information about a population, which necessitates their specific treatment during EDA (Exploratory Data Analysis).

Furthermore, errors in data collection and processing might result in "error-outliers." These metrics are frequently not representative of the group we are interested in, necessitating treatment. In our research model, we do not treat the outliers as we deal with the numerical values of the air components. Since the AQI dataset can contain extreme values of daily air constituents, we choose not to treat the outliers to maintain the reliability of the machine learning model.

3.5 Feature Extraction/Selection:

The process of extracting the most consistent, non-redundant, and relevant features to employ in model creation is known as feature selection. As the number and variety of datasets grow, it's more critical than ever to reduce them methodically. The fundamental purpose of feature selection is to improve the predictive model's performance while lowering the modelling cost.

For data scientists, feature selection is a useful asset. Understanding how to choose important characteristics in machine learning is critical to the algorithm's effectiveness. Irrelevant, redundant, and noisy features can clog up a learning system, lowering performance, accuracy, and computing cost. As the amount and complexity of the average dataset grow rapidly, feature selection becomes increasingly crucial. In this prediction we have excluded a few constituents that make up the AQI, the reason being the availability of data. The data for the constituents 'NO', 'NOx', 'Benzene', 'Toluene', 'Xylene' is not monitored due to lack of efficiency.

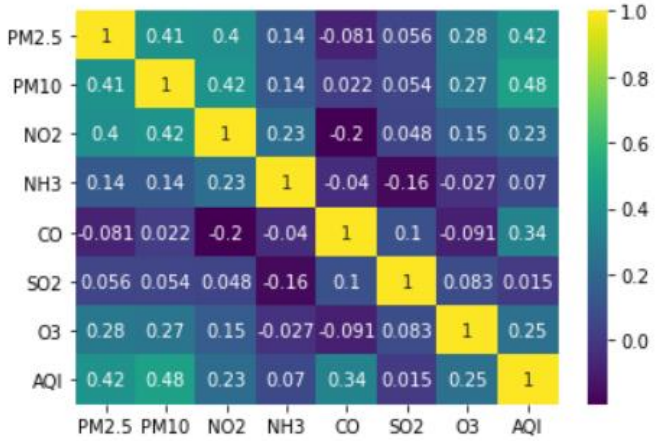

Fig 6: Correlation matrix of the constituents.

4. Applying Models:

Machine Learning Regression Algorithms

Testing the model with 70-30 split and applying various Regression Algorithms to check the accuracy and use the best model for prediction based on performance evaluation metrics.

4.1 Linear Regression

Linear regression is the first regression model used to predict AQI levels. It's commonly used for predictive analysis and as a baseline for performance comparisons.

Linear regression is used for finding the relationship between

1. Independent variable called predictor.
2. Dependent variable known as a response.

Y = B0 + Bi*Xi is a formula for linear regression.

Where Y is the dependent variable.
Bi – Weights for 'i' features
Bi – Weights for 'i' features
Xi – 'i' independent variable

4.2 Decision Tree Regression

Decision tree regression examines an object's properties and trains a model in the form of a tree to forecast data in the future to provide meaningful continuous output. The output/results are not discrete, i.e., they are not represented exclusively by a discrete, known set of numbers or values.

A decision tree is constructed from the top-down, starting with a root node, and involves partitioning the data into subsets containing instances with comparable values (homogenous). The homogeneity of a numerical sample is calculated using standard deviation. The standard deviation of a numerical sample that is perfectly homogeneous is 0.

4.3 Polynomial Regression

Polynomial Regression is a regression approach that uses an nth degree polynomial to represent the connection between a dependent(y) and independent variable(x). The equation for polynomial regression is as follows:

$$Y = b0+b1x1+ b2x12+ b2x13+..... bnx1n$$

Machine learning, it's also known as the specific case of Multiple Linear Regression. Because we turn the Multiple Linear regression equation into Polynomial Regression by adding certain polynomial terms.
It's a linear model that's been tweaked a little to improve accuracy. The training dataset for polynomial regression is non-linear in character. To fit the intricate and non-linear functions and datasets, it employs a linear regression model.

4.4 XGBoost

Execution speed and model performance are the two key reasons to employ XGBoost.
On classification and regression predictive modelling issues, XGBoost dominates structured or tabular datasets. It is the go-to method for competition winners on the Kaggle competitive data science platform, according to the evidence.
The beauty of this powerful algorithm is its scalability, which allows for rapid learning via parallel and distributed computing while still utilizing memory efficiently.
XGBoost is a method of ensemble learning. It may not always be enough to rely on the outcomes of a single machine learning model. Ensemble learning is a method for combining the predictive abilities of numerous learners systematically. The result is a single model that combines the outputs of numerous models.

4.5 Random Forest

A Random Forest is an ensemble technique that uses several decision trees and a technique called Bootstrap and Aggregation, sometimes known as bagging, to solve both regression and classification problems. Instead of depending on individual decision trees, the main idea is to aggregate numerous decision trees to determine the outcome.
As a fundamental learning model, Random Forest uses several decision trees. Row and feature sampling are done at random from the dataset, resulting in sample datasets for each model. This part is called Bootstrap.

- Create a machine learning model.
- Set the baseline model that you want to achieve
- Train the data
- Provide an insight into the model with test data
- Compare performance metrics
- Try improvising the model if the results are not satisfactory.
- Interpret the results.

4.6 Performance Metrics

The performance of the models mentioned in this project is evaluated using the following performance measures.

1. The root means square error (RMSE) is an aggregate of the difference between expected and

actual values over numerous observations. The less the RMSE, the better.

2. The average difference between the expected and actual values is represented by the Mean Absolute Error (MAE). The smaller the MAE, the better.
3. The R2 coefficient of determination is used to determine how near the anticipated data points are to the regression line that has been fitted. It's a metric for how well something fits. Better model performance is indicated by a higher R2 value.

5. Model Evaluation and Results

RandomForest Regressor

```
RFreg_model = RandomForestRegressor()
RFreg_model.fit(X_train,y_train)
prediction2 = RFreg_model.predict(X_test)
rmse_RFreg = np.sqrt(mean_squared_error(y_test, prediction2))
print('RMSE value is = {}'.format(rmse_RFreg))
r2_RFreg = r2_score(y_test, prediction2)
print('R-squared value is {}'.format(r2_RFreg))
from sklearn.metrics import mean_absolute_error
rmse_rfreg1=mean_absolute_error(y_test,prediction2)
print('R-squared value is {}'.format(rmse_rfreg1))
```

```
RMSE value is = 23.61491553711942
R-squared value is 0.6339054735044173
R-squared value is 14.280586122036702
```

Fig 7: Fitting the Regression Model.

Random Forest Regressor is chosen based on the performance metrics and further prediction is carried out. A data frame is created to show the differences between the actual value and the predicted value of AQI.

Based on India AQ Standards we have classified the AQIs as per the category.

```
print("Accuracy:",count/len(final)*100)
```

```
Accuracy: 79.1044776119403
```

Fig 8: Accuracy of the Model.

The model achieves an accuracy of 79.10%.

A scatter plot is plotted between the actual and predicted values.

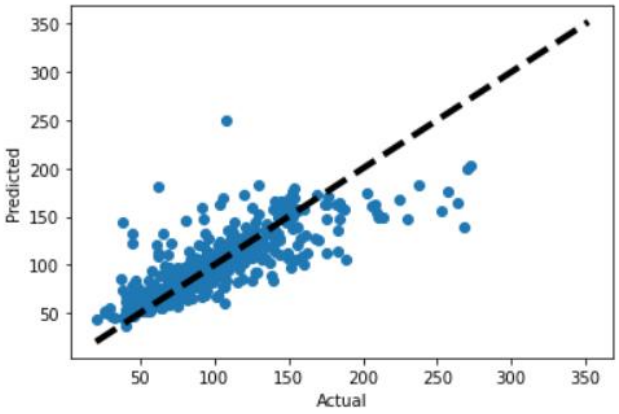

Fig 9: Scatter plot of actual and predicted values.

The relationship between the actual and predicted values are positive. Hence, we can conclude that our prediction is accurate.

With the diminishing quality of the AQI, we have analyzed, predicted and suggested possible solutions for the betterment of the future.

6. Conclusion

- **3R's**

Reduce, Recycle, and Reuse are the three R's of environmental control.

Recycling and reuse are beneficial to the environment since they not only help to save resources and use them wisely, but they also help to minimise pollution emissions. Furthermore, recycled materials require less energy to produce.

Planting and growing as many trees as possible are referred to as afforestation. Planting trees provides a number of environmental advantages and aids in the release of oxygen.

- **Making use of public transportation:**

Using public transit, which uses less gasoline and electricity, is a surefire way to minimise air pollution; even carpools can assist. Taking public transportation can save you money while also reducing the amount of gasoline and gas that is discharged into the atmosphere.

- **Electric vehicles:**

Electric vehicles have a significant advantage in terms of improving the region's air quality. It aids in the

reduction of CO2 emissions, particulate matter (PM), nitrogen oxide (NOx), carbon monoxide (CO), and other pollutants.

7. Acknowledgement

We would like to express the deepest gratitude to our guide Prof. Mr. Santhosh Rebello who has been a constant supporter. He continually and convincingly conveyed a spirit of adventure concerning research and scholarship and excitement concerning teaching. Without his guidance and persistent help, this paper would not have been possible.

8. Bibliography

[1] Machine Learning-Based Prediction of Air Quality-Yun-Chia Liang, Yona Maimury, Angela Hsiang-Ling Chen, and Josue Rodolfo Cuevas Juarez.

[2] Air Quality Analysis and Prediction-Surender Sampath, University of London

[3] National Air Quality Index Report-Central Pollution Control Board (cpcb) 2014-15

[4] Indian Air Quality Prediction and Analysis using Machine Learning-Mrs. A. Gnana Soundari, Mrs.J. Gnana Jeslin, Akshaya A.C

[5] Study of Ambient Air Quality Trends and Analysis of Contributing Factors in Bengaluru, India-Amrita Thakur.

[6] Democratizing Digital Solutions to Improve Public Health and Urban Air Quality-Microsoft Urban Futures-Summer 2020.

[7] Status quo analysis of various segments of electric mobility and low carbon passenger road transport in India – NITI Aayog

[8] Central Pollution Control Board www.cpcb.nic.in

[9] https://app.cpcbccr.com/AQI_India/

[10] www.kaggle.com

[11] www.analyticsvidhya.com