International Journal of Science and Research (IJSR)
ISSN: 2319-7064
SJIF (2022): 7.942

Determining the Air Quality Index of Indian Cities Using Machine Learning

Rishit Garg

Lotus Valley International School
rishg07[at]gmail.com

Abstract: ***In this paper, I have applied the concepts of machine learning (ML) and regression analysis for determining the Air Quality Index (AQI) of various Indian cities during the period of 2015 to 2020. The article starts with a brief introduction to machine learning and its types, and then discusses the basic steps to train an ML model – missing values treatment, encoding, model training and testing. The dataset used in this paper has more than 10 attributes and the target variable – the AQI score. These include $PM_{2.5}$, PM_{10}, NO, NO_2, NH3, CO, SO_2, O_3, Benzene, Toluene and Xylene. The paper also highlights the impact of these attributes on the health of human beings. The XG Boost regressor model was trained using the dataset and 5 different random states were considered for testing the model's consistency. An average R-squared value of 0.9318was calculated for the trained model.***

Keywords: Machine learning, Regression, Air quality, XGBoost

1. Introduction

Machine learning (ML) is not a new concept and was coined by Arthur Samuel, a computer scientist at IBM in 1959. It is not a universal term and has been defined differently by different people. Some of the common definitions are [1]:

- Arthur Samuel defined machine learning as "the field of study that gives computers the ability to learn without being explicitly programmed."
- Ethem Alpaydin in his textbook defined machine learning as the field of "Programming computers to optimize a performance criterion using example data or past experience."

ML is not interchangeable with artificial intelligence (AI), but it is a subset of AI. AI is an intelligent system which can perform various complex jobs while machine learning is about training a machine to perform a job – in a way like we train a human brain. In ML, the algorithm is given access to data and is allowed to improve themselves with minimum human intervention [2]. Thus, both AI and ML are the best in their respective fields.

1.1. Types of Machine learning

There are basically three main types of machine learning [3].

1.1.1. Supervised ML

In this type of machine learning, the machine learning algorithms are fed historical input and output data. The algorithm then computes relationship between each input and output variables allowing the algorithm to train the model to predict outputs as closely aligned with the desired results as possible. The fed data is 'labelled' and requires human labour to make it understandable for the computer.

1.1.2. Unsupervised ML

In this type, the algorithm is fed with 'unlabelled' data. This reduces the human labour required to make the datasheet readable, thus allowing it to work with much larger datasets. The computer creates a hidden structure as it works with unlabelled data. This creation of hidden structures makes it more versatile than supervised learning. The algorithm can adapt to the data by changing its hidden structure.

1.1.3. Reinforcement ML

It is the closest form of learning to human nature as it improves itself through a trial-and-error method. The algorithm learns by interacting with its environment and getting a positive or negative reward. In every iteration of the algorithm, the output result is given to the interpreter which decides whether it is a favourable or unfavourable outcome. If the outcome is the correct solution, then the interpreter gives the algorithm reinforcement by giving it a reward. If the outcome found is the incorrect solution, then the interpreter gives the algorithm a punishment until it gets the correct solution. Thus, this improves the algorithm's accuracy the more it is trained.

1.2. Air Quality Index

This paper uses the Air Quality Index (AQI) dataset for analysis. AQI may be defined as a single number for reporting the air quality with respect to its effects on human health[4]. Air pollution contributes to increased mortality and hospital admissions. It can also cause birth defects, serious developmental delays in children, and reduced activity of the immune system, leading to a number of diseases [5].Certain important parameters that contribute significantly to the air quality of a particular area are discussed below.

1.2.1. PM 2.5:Particulate matter 2.5 ($PM_{2.5}$), refers to the tiny droplets in the air that are two- and one-half microns or less in width. Exposure to it can cause short-term health effects such as eye, nose, throat and lung irritation, coughing, sneezing, runny nose and shortness of breath. Exposure to fine particles can also affect lung function and worsen medical conditions such as asthma and heart disease [6].

1.2.2. NO_2: Inhalation of nitrogen dioxide by children increases their risk of respiratory infection and may lead to poorer lung function in later life. Nitrogen dioxide can decrease the lungs' defences against

bacteria making them more susceptible to infections. It can also aggravate asthma [7].

1.2.3. NH3: Exposure to high concentrations of ammonia in air causes immediate burning of the nose, throat and respiratory tract. This can cause bronchial and alveolar edema, and airway destruction resulting in respiratory distress or failure. Inhalation of lower concentrations can cause coughing, and nose and throat irritation [8].

1.2.4. CO: Carbon monoxide is harmful because it binds to haemoglobin in the blood, reducing the ability of blood to carry oxygen. This interferes with oxygen delivery to the body's organs. The most common effects of CO exposure are fatigue, headaches, confusion, and dizziness due to inadequate oxygen delivery to the brain [9].

1.2.5. SO2: Sulfur dioxide affects the respiratory system, particularly lung function, and can irritate the eyes. Sulfur dioxide irritates the respiratory tract and increases the risk of tract infections. It causes coughing, mucus secretion and aggravates conditions such as asthma and chronic bronchitis [10].

1.2.6. Aromatic organic compounds: Benzene and ethylbenzene exposure is linked with an increased risk of leukaemia and hematopoietic cancers. Toluene and xylene are non-carcinogenic, but they may produce reproductive adverse effects; especially when exposures are chronic at low to high concentrations [11].

1.2.7. O3: Breathing ground-level ozone can trigger a variety of health problems including chest pain, coughing, throat irritation, and congestion. It can worsen bronchitis, emphysema, and asthma. Ozone also can reduce lung function and inflame the lining of the lungs. Repeated exposure may permanently scar lung tissue [12].

This paper uses AQI dataset of 26 cities across India during 2015-2020 with more than 10 attributes.XGBoost regressor model is used to train and predict the AQI score. The article also highlights certain key concepts of machine learning and focuses on an important environmental aspect: Air quality.

2. Method

2.1. Data and tools

The data for the model has been collected from a public platform, Kaggle on[13].Google colab has been used as the platform for coding using python language. The following sections provide detailed insights on regression analysis and XGBoost regressor model used in the paper.

2.1.1. Regression Analysis

Regression is a mathematical way of determining the relationship between a dependent variable and other independent variables [14]. To conduct regression analysis:

a) We gather all the data and mark it as points on a graph.
b) Derive the best fit line (called the regression line) with the minimum sum of squared errors using the model (see figure 1).
c) This regression line is the best explanation of the relationship between the independent and dependent variables.

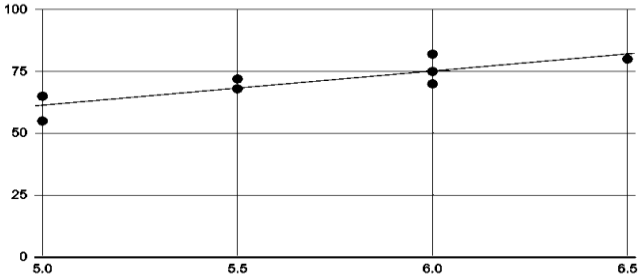

Figure 1: Scatter plot showing various data points and the regression line

R-square is a statistical measure which is used in a regression model and is the "percent of variance explained" by the model. That is, R-squared is the fraction by which the variance of the errors is less than the variance of the dependent variable[15]. Thus, it shows how well the data fits in with the model.

$$R - Squared = \frac{SS_{regres\ sion}}{SS_{total}}$$

Where $SS_{regression}$ is the sum of squares due to regression and measures how well the regression model represents the data used for modelling and SS_{total} is the total sum of squares and measures the variation in the observed data.

2.1.2. XGBoost

XGBoost is a machine learning tool using multiple techniques: 1) Regularization, 2) Gradient Boost, 3) Unique Regression tree etc. The XGboost builds trees using Regression or Classification techniques. It was designed to work with large and complicated data sets. It builds trees using the similarity scores and then calculates the output values of the leaves.It was designed by Tianqi Chen as part of the DMLC group [16].

3. Results and Discussion

The flowchart in Figure 2 shows the steps that were followed and are further detailed in next section along with the codes.

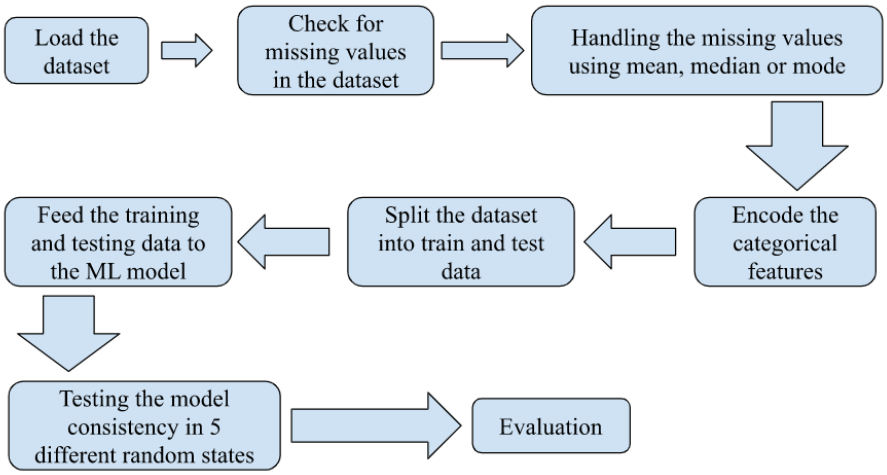

Figure 2: Flowchart of the steps

The original data looked as shown in Figure 3. There were 29,531 rows and 16 columns. After obtaining the data, .info()was used to view the data types of the features (float data type for numerical columns and object data type for categorical/non-numerical columns). It was found that there were 13 numerical features and only 3 categorical features namely: city, date and AQI Bucket (see Figure 4).Proceeding as per the steps mentioned in Figure 2, the next step is to find the missing values for all the features, both numerical and categorical.

```
# first 5 rows of the dataframe
city_day.head()
```

	City	Date	PM2.5	PM10	NO	NO2	NOx	NH3	CO	SO2	O3	Benzene	Toluene	Xylene	AQI	AQI_Bucket
0	Ahmedabad	2015-01-01	NaN	NaN	0.92	18.22	17.15	NaN	0.92	27.64	133.36	0.00	0.02	0.00	NaN	NaN
1	Ahmedabad	2015-01-02	NaN	NaN	0.97	15.69	16.46	NaN	0.97	24.55	34.06	3.68	5.50	3.77	NaN	NaN
2	Ahmedabad	2015-01-03	NaN	NaN	17.40	19.30	29.70	NaN	17.40	29.07	30.70	6.80	16.40	2.25	NaN	NaN
3	Ahmedabad	2015-01-04	NaN	NaN	1.70	18.48	17.97	NaN	1.70	18.59	36.08	4.43	10.14	1.00	NaN	NaN
4	Ahmedabad	2015-01-05	NaN	NaN	22.10	21.42	37.76	NaN	22.10	39.33	39.31	7.01	18.89	2.78	NaN	NaN

Figure 3: Output showing the first five rows of the dataset and 16 features

```
#getting some information about the dataset
city_day.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 29531 entries, 0 to 29530
Data columns (total 16 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   City        29531 non-null  object
 1   Date        29531 non-null  object
 2   PM2.5       24933 non-null  float64
 3   PM10        18391 non-null  float64
 4   NO          25949 non-null  float64
 5   NO2         25946 non-null  float64
 6   NOx         25346 non-null  float64
 7   NH3         19203 non-null  float64
 8   CO          27472 non-null  float64
 9   SO2         25677 non-null  float64
 10  O3          25509 non-null  float64
 11  Benzene     23908 non-null  float64
 12  Toluene     21490 non-null  float64
 13  Xylene      11422 non-null  float64
 14  AQI         24850 non-null  float64
 15  AQI_Bucket  24850 non-null  object
dtypes: float64(13), object(3)
memory usage: 3.6+ MB
```

Figure 4: Output showing the data type of 16 features (columns) of the dataset

If more than 30% of the data is missing in a particular column, then ideally that column should be dropped. In our dataset having 29,531 rows, the columns PM_{10}, NH_3 and Xylene had more than 30% of data values missing. However, as discussed in section 1.2 above, we cannot drop these columns, since they impact the AQI index. So, we will find the data distribution and skewness of these columns while handling the missing values. Figure 5 shows the skewness of the PM2.5 column which is right-skewed. Out of mean, median and mode, median nullifies the skewness of data to the largest extent. So, we use median to fill the missing values in all the numerical columns. For the categorical column of AQI bucket, mode is used to fill the missing values. The city column value has been used to find the mode as the mode may depend on the AQI Bucket of each individual city based on geographical location and climate.

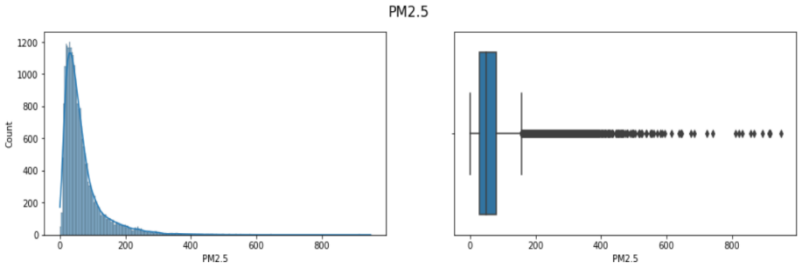

Figure 5: Distribution plot of PM2.5 feature of the dataset along with box plot showing right-skewness

(a) Before missing values treatment

```
# checking for missing values
city_day.isnull().sum()
```

```
City              0
Date              0
PM2.5          4598
PM10          11140
NO             3582
NO2            3585
NOx            4185
NH3           10328
CO             2059
SO2            3854
O3             4022
Benzene        5623
Toluene        8041
Xylene        18109
AQI            4681
AQI_Bucket     4681
dtype: int64
```

(b) After missing values treatment

```
# After Missing Values Treatment
# checking for missing values
city_day.isnull().sum()
```

```
City          0
Date          0
PM2.5         0
PM10          0
NO            0
NO2           0
NOx           0
NH3           0
CO            0
SO2           0
O3            0
Benzene       0
Toluene       0
Xylene        0
AQI           0
AQI_Bucket    0
dtype: int64
```

Figure 6: a) Output showing sum of missing values of each column in original dataset. b) Output showing 0 missing values in all columns after missing value treatment

As shown in Figure 6(b), all missing values of the data have been filled successfully. Now, as per the flowchart shown in figure 2, the next step is to encode the data in the categorical features. Label encoding is used for this and numbers are assigned to the non-numerical data. This is done as the model can only interpret numerical data.Once encoding is done, the model has to be trained using the processed data. Both the target and features are split into training data and testing data using the train_test_split function (see Figure 7). The training data is used to train our model while the testing data is used to test the model to calculate its accuracy in predicting the AQI.

```
X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size=0.2, random_state=2)
```

```
print(X.shape, X_train.shape, X_test.shape)
```

```
(29531, 15) (23624, 15) (5907, 15)
```

Figure 7: Output showing use of train_test_split function to split the data into training and testing data in 80:20 proportion (test_size = 0.2) with random state 2.

Assigning a particular random state allows us to get the same train and test data across executions and reproduce the same train test split each time you run the code. Checking the R-squared value for various random states allows us to verify the consistency of the model. As observed in Table 1, the R-squared values for both train and test data across various random states do not differ significantly and hence prove that the model was trained properly.

Table 1: R-squared values calculated for training data and testing data split at different random states

Random State	R-squared value (train)	R-squared value (test)
1	0.9314	0.9247
2	0.9305	0.9148
3	0.9331	0.9093
4	0.9313	0.9198
5	0.9326	0.9118
Mean	0.9318	0.9161
Std. Dev	0.0011	0.0062

The model shows an average R-squared value of 0.9318 with only 0.11% standard deviation. This acute closeness of R-squared value to 1 shows that the model was trained suitably using the dataset and can be deployed for future predictions.

4. Conclusion

In this paper an ML model was used to predict the AQI of 26 Indian cities during the period from 2015 to 2020. The XGBoost model was trained using the AQI dataset collected from Kaggle following the basic steps of ML modelling. The

pre-processing of data involving searching for missing values, treating them, encoding of categorical features, among others, were explained in detail and executed in the paper. The model was run through various random states and the R-squared value in each case was found and compared. An average R-squared value of the trained model was reported as 0.9318 with a minor standard deviation of 0.11%. Besides, the concepts of ML model execution and data processing, the paper also discusses the implications of various features of AQI on the physical wellbeing of people.

References

[1] ElNaqa, I. and Murphy, M.J., 2015. What is machine learning?. In machine learning in radiation oncology (pp. 3-11). Springer, Cham.
[2] Attaran, M. and Deb, P., 2018. Machine learning: the new'bigthing'for competitive advantage. International Journal of Knowledge Engineering and Data Mining, 5(4), pp.277-305.
[3] Morgan, D. and Jacobs, R. (2020) 'Opportunities and Challenges for Machine Learning in Materials Science', *Annual Review of Materials Research*, 50(1). doi:10.1146/annurev-matsci-070218-010015.
[4] Bishoi, B., Prakash, A. and Jain, V.K. (2009). A Comparative Study of Air Quality Index Based on Factor Analysis and US-EPA Methods for an Urban Environment. Aerosol and Air Quality Research, 9(1), pp.1–17. doi:10.4209/aaqr.2008.02.0007.
[5] Kampa, M. and Castanas, E. (2008). Human Health Effects of Air Pollution. Environmental Pollution, [online] 151(2), pp.362–367. doi:10.1016/j.envpol.2007.06.012.
[6] www.health.ny.gov. (2018). Fine Particles (PM2.5) Questions and Answers. [online] Available at: https://www.health.ny.gov/environmental/indoors/air/pmq_a.htm#:~:text=Exposure%20to%20fine%20particles%20can,as%20asthma%20and%20heart%20disease.
[7] Ministry for the Environment. (2021). Nitrogen dioxide. [online] Available at: https://environment.govt.nz/facts-and-science/air/air-pollutants/nitrogen-dioxide-effects-health/#:~:text=Effects%20on%20health.
[8] New York State Department of Health (2019). The Facts About Ammonia. [online] Ny.gov. Available at: https://www.health.ny.gov/environmental/emergency/chemical_terrorism/ammonia_tech.htm.
[9] ww2.arb.ca.gov. (n.d.). Carbon Monoxide & Health | California Air Resources Board. [online] Available at: https://ww2.arb.ca.gov/resources/carbon-monoxide-and-health#:~:text=Carbon%20monoxide%20is%20harmful%20because.
[10] Queensland;, c=AU; o=The S. of (n.d.). Sulfur dioxide | Air pollutants. [online] www.qld.gov.au. Available at: https://www.qld.gov.au/environment/management/monitoring/air/air-pollution/pollutants/sulfur-dioxide#:~:text=Sulfur%20dioxide%20affects%20the%20respiratory.
[11] Masekameni, M., Moolla, R., Gulumian, M. and Brouwer, D. (2018). Risk Assessment of Benzene, Toluene, Ethyl Benzene, and Xylene Concentrations from the Combustion of Coal in a Controlled Laboratory Environment. International Journal of Environmental Research and Public Health, [online] 16(1), p.95. doi:10.3390/ijerph16010095.
[12] www.iowadnr.gov. (n.d.). Effects of Ground Level Ozone. [online] Available at: https://www.iowadnr.gov/Environmental-Protection/Air-Quality/Air-Pollutants/Effects-Ozone#:~:text=Breathing%20ground%2Dlevel%20ozone%20can.
[13] https://www.kaggle.com/datasets/rohanrao/air-quality-data-in-india.
[14] Gallo, A. (2015). A Refresher on Regression Analysis. [online] Harvard Business Review. Available at: https://hbr.org/2015/11/a-refresher-on-regression-analysis.
[15] Nau, R. (2019). What's a good value for R-squared? [online] Duke.edu. Available at: https://people.duke.edu/~rnau/rsquared.htm.
[16] madrasresearchorg (2021). XGBOOST. [online] MSRF|NGO. Available at: https://www.madrasresearch.org/post/xgboost.