WWW.IJRPR.COM

**International Journal of Research Publication and Reviews**

Journal homepage: www.ijrpr.com ISSN 2582-7421

# Air Quality Index Analysis Using Machine Learning

***Vinay Raj A S[1], P Rachitha[2], Priya S N[3], Rahul[4], Bharathi S[5]***

*[1]Assistant Professor, Department of Computer Science and Engineering, MVJ College of Engineering, Bangalore, Karnataka, India.*
*[2,3,4,5]Undergraduate Scholar, Department of Computer Science and Engineering, MVJ College of Engineering, Bangalore, Karnataka, India.*

**A B S T R A C T**

When it comes to the spread of diseases like lung cancer, autism, asthma, low birthweight, and others, air pollution is a significant environmental risk factor. To protect the health and welfare of its citizens, governments in emerging nations must play a significant role in regulating air quality.

Localized variations in air pollution result from a variety of pollutant sources, including combustion of fossil fuels, excessive traffic congestion, industrial pollutants, and atmospheric conditions.

Today, practically all industrial and metropolitan regions require the government to analyse and preserve the quality of the air, making it one of their most important tasks. The concentrations of air pollutants such SO2, NO, PM2.5, O3, and PM10 are examined in this study using machine learning methods.

This model assesses the air quality based on various pollutant concentrations to efficiently extract characteristics and make judgements. A machine learning model is developed to anticipate the air quality index based on past air quality data using linear regression and the SARIMA model. The trials' findings show that the proposed model may be used to monitor air quality and predict its level in the future. The model's performance on the train data is 71.69%.

**Keywords:** Air Pollution, air pollutants, air quality index.

## 1. Introduction

Recent invention has accelerated the pace of development. This would lead to a range of seriousenvironmental issues, such as air pollution, sound pollution, deforestation, water pollution, acidrain, the manufacturing of toxic compounds, and others when coupled with a dramatic increasein the number of people and automobiles. Industrialization has greatly risen to keep up with the demands of a growing population. This might result in the release of harmful gases into theenvironment from a number of industries, which could have a significant impact on air pollution in urban areas throughout the world. This demonstrates that due to the high concentration of hazardous gases and other airborne particles that have a detrimental influence on people's health, the air that people are inhaling is not clean but rather contaminated. Pollution causes a reduction in air quality.

Air pollution is a major problem in the majority of urban places. Residents need to pay attentionto the air they are breathing. The National Ambient Air Monitoring Network produces data thatillustrates the various air pollution concentrations at various levels, however this data is challenging for the common individual to interpret. As a result, India's cities' national Air Quality Index (AQI) is created by the Central Pollution Control Board (CPCB). Information about a location's degree of air pollution is available from the air quality index (AQI). This demonstrates that the AQI connects to a variety of detrimental health impacts and assesses theactual air quality around us in a meaningful way.

According to the CPCB, the AQI (Benzene) will be calculated using 12 air pollutants, includingNO2, SO2, CO, O3, PM10, PM2.5, NH3, toluene, and xylene. The criterion pollutants (i.e., PM10, PM2.5, SO2, NO2, CO, and O3) are usually used to calculate the AQI, even though it may be challenging to employ several of the pollutants from the list of 12 pollutants. The AQIobjectives, the average period, the data available, the frequency of monitoring, and themeasurement methods, on the other hand, are dependent on the removal of pollutants. The AQIis a figure that government agencies use to assess air pollution levels and educate the public, according to a straightforward definition

## 2. Literature Review

K. Mahesh Babu and J. Rene Beulah [1] Data preparation and purification, record-missing detection, thorough review, model creation, and model evaluation were the first steps in the analytical process. When compared to classification records on the public test set, the decision tree approach procedure exhibits exceptional accuracy. This approach can assist India's urban regions in forecasting the future of air quality and its reputation based on their capacity to respond.

Khalid M. O. Nahar and Mohammad Ashraf Ottom both [2] The Air Quality Index (AQI) in the various nations, each of which utilizes a distinct management plan, must be recognised in order to calculate the amount of pollutants and pollution. The main polluting cause that contributes to the

impurity of the air is specified in the classification of the AQ as a polluted zone. This model (DT) uses machine learning methods including decision trees and logistic regression. It is possible to predict the polluting component using this model and the daily measurements.

Gaurav Shilimkar, Shivam Pisal, and Ritik Sharma [3] The Python machine learning methodsapplied in this model were tested using Jupiter notebook.

At the outset of the model, the key elements that are taken into account for the outcome are selected. Since these features are related, the model can be trained using them. The air quality list is predicted by this model employing original computations including direct relapse, Decision Tree, and SARIMA Model. The results show that neither the Decision Tree nor the SARIMA Model computation can forecast air quality with greater accuracy

Avnish Bora and Laxmi Chaudhary both [4] For model training, this model gathers the requireddata from modern IOT devices using a variety of sensors and devices. The development of cities in the future will necessitate a greater knowledge of air quality issues, which may be researched and assessed utilizing machine learning-based AI algorithm models. These new innovations, such portable electronics and inexpensive sensors, effectively provide the model with vast amounts of data. It helps to reduce air pollution, which reduces the likelihood that individuals may have health issues. It also helps make other species in our surroundings, suchplants and animals, live longer.

Both Heniel Kashyap and Udit Ranjan Kalita [5] The biggest problem that many developing nations have is air pollution. Pollution is now becoming worse and worse every day for a variety of reasons, including industry, population increase, the development of new technology,the chemical industry, and many other sectors. This technique takes use of semiconductor sensors, such as the MQ9 and MQ7, which can anticipate the AQI and identify some air pollution-causing agents.

## 3. PROPOSED SYSTEM

The goal of the suggested approach is to raise awareness of the Air Quality Index (AQI), a statistic used to assess air quality, among the general public. Whether or if the environment's living things, including people, are at risk from the air.The process of developing the model starts with gathering the appropriate data. The Central Pollution Control Board (CPCB) websiteis where the dataset was obtained. Pre-processing the gathered data is necessary after acquiring the necessary data. The process of feature selection eliminates unnecessary records while taking into account the information required to forecast the AQI. Also, the dataset is shown.

Next, a data cleaning process was carried out because the model has to be more accurate. To improve the model's score during this phase of missing value treatment, the outlier is checked. The model is then trained using machine learning methods like linear regression and the SARIMA Model using cleaned train data, and test data is used to calculate the model's accuracy score. In addition to showing the user the category that the AQI value belongs to, the model also displays the AQI value. Because the model needs to be more accurate, the data was then cleaned. At this stage of the missing value treatment, outliers are examined in order to increasethe model's accuracy. The model is then trained using machine learning methods like linear regression and the SARIMA Model using cleaned train data, and its accuracy is tested using test data.

## 4. IMPLEMENTATION AND PROCESS

### 1. Dataset :

The dataset that is available on the CPCB website contains over 30,000 data tuples with more than 16 characteristics. The increased dimensionality has a detrimental effect on the model construction. Among the characteristics were forecasts for the pollution levels of 4 distinct pollutants. Therefore, we diluted all pollutants to NO2 levels. The same approach may be usedto other pollutant qualities depending on the circumstance. Additionally, the data spans six years and includes hourly figures for each day. The parameters used in the data collection (O3) comprised particulate matter (PM2.5 and PM10), nitrogen monoxide (NO), nitrogen dioxide (NO2), nitrogen oxide (NOx), ammonia (NH3), carbon monoxide (CO), Sulphur dioxide (SO2), and ozone.

### 2. Data Pre-Processing :

Use the AQI, PM10, and CO measures to analyse the data and plot the most polluted cities. Data pre-processing is a data mining technique used to transform the raw data into an effectiveand usable shape. According to the results for the AQI, PM10, and CO measures, the following figure shows the cities with the highest levels of pollution in the dataset.

### 3. Attribute Selection :

The new characteristic is selected from the set of attributes that are readily available. The elements that significantly contribute to air pollution and have the greatest value row-wise make up the Air Quality Index.

### 4. Standard Scaler :

By using the Standard Scaler tool, one may scale the value distribution in machine learning so that the mean and standard deviation of the observed data are both 0. On the basis of this model,I'll show how to use Standard Scaler in machine learning. The functional range of the input dataset is normalized using Standard Scaler, a crucial tool that is frequently used as a pre- processing step before many machine learning models. Standard Scaler comes into play when the input dataset's characteristics vary noticeably across their respective ranges or are just measured in various units of measurement.

After using Standard Scaler to remove the mean, the data are scaled to the unit variance. However, outliers have an impact that narrows the range of characteristic values whencalculating the empirical mean and standard deviation. These changes in the initial characteristics may cause problems for a number of machine learning models. If one of the qualities in a model that calculates distance, for example, has a wide range of values, that particular characteristic will determine how far something is.

The Standard Scaler is founded on the idea that variables measured at different scales don't all contribute equally to the model's fit and learning function, and they could even introduce bias.We must thus normalise the data ( $\mu = 0$, $\sigma = 1$) that is frequently used to solve this possible problem before including it into the machine learning model. Characteristics are made uniformby getting rid of the mean and scaling to one variance. The formula for calculating a sample's standard score is $z = (x - u) / s$.

**5. Data Cleaning :**

Data cleansing is a vital stage in the data preparation process. If the raw data is not cleaned,the results of the model may suffer, and the model may not be suitable for the dataset. Manual data cleansing is insufficient, therefore I use the strategies listed below: For better outcomes, look for missing values and use the median function to fill in any empty numbers.Find any duplicate values, then get rid of them.

**6. Data Visualization :**

For both small- and large-scale data presentations, data visualisation, or the graphic representation of information and data, is crucial. Using data visualisation tools, which includegraphic components like charts, graphs, and maps, is a comprehensible approach to spot trends,outliers, and patterns in data.

**7. Linear Regression**

Probabilistically, linear regression was where most academicians first experimented with machine learning. The core principle guiding its operation is the fitting of one or more. Typically, n-variable datasets are represented as an n-dimensional line with independent variables and the dependent variable. A maximum number of mistakes would be avoided, according to the line's designers, if all the occurrences could fit within it. By adjusting the model's parameters, linear regression is possible to continuously learn under machine learning.

These parameters include x0, x1, x2,..., xp. The method most usually employed for optimization is called gradient descent. All parameters are updated by subtracting the previous value from the derivative multiplied by a predetermined learning rate. It works by partially calculating the loss function. The easiest way to change the learning rate is by trial and error, although more complicated methods, such meta-heuristics, can also be utilized. Another variable that may still be changed is how much generality the model introduces. Reducing the possibility of overfitting and enhancing the model's resilience are key components of regularization. In linear regression, two regularization techniques are used: Lazo and ridge regression. When a feature's coefficient is set to zero, lasso regularization will maintain any significant features and discard any less-significant ones. The goal of ridge regularization, in contrast, is not to eliminate a feature but to lower the size of the coefficients to obtain a lesservariance in the model.

**8. SARIMA Model**

Temporary SARIMA ARIMA, or, to be more precise, ARIMA with a seasonal component. As mentioned earlier, the ARIMA statistical analysis model uses time-series data to either better understand the data set or predict future patterns. The modelling of time series data is a highly personalised and subjective process. Different parameters can be used for the same time series. As a result, there is no conclusive solution. The best solution is the one that solves the demandsof the company in an efficient manner. It might be difficult to fully understand the model- building process due to the amount of subjectivity involved. I was able to structure the findingsinto a framework after doing several investigations, presentations, and implementations.

The mean and variance are consistent throughout the data. As a result, the data do not need tobe changed. Currently, we are examining the information's seasonal and trend components. After the model was successfully equipped with the data, it is necessary to look at the residual plots to confirm the model's validity. An efficient forecasting technique will result in residualsthat have the qualities listed below: There are some uncorrelated residuals. If there are correlations between the residuals, the residuals still have data that can be used to make forecasts. The mean of the residuals is zero. The forecasts are inaccurate if the residuals have a mean that is not zero.

## 5. RESULT :-

This model is developed using common machine learning techniques like linear regression andthe SARIMA Model (IDE) with the Python PyCharm Integrated Development Environment. mainly because the quantity of airborne contaminants impacts the value of the air quality index. If the value of the feature changes concurrently with the value of the independent variable, thenfeatures and independent variables are likely related.

The model's efficacy (RMSE) may be evaluated using a number of evaluation measures, including Mean Square Error (MSE), R- Square Error, and Root Mean Square Error. The data utilised for the trials is split into train and test data, with the model considering 80% of the training data and 20% of the test data, which is used to determine whether the model is functioning properly.

## 6. CONCLUSION:-

With the help of the Air Quality Index (AQI), which determines the amount that the air that webreathe is contaminated, this study aims to fully understand the AQI. Understanding the AQI is important because people won't be as concerned about it and less likely to take action to lessen it if they are unaware of the worst effects or hazards connected with air pollution. If theyare able to take action, this plan can assist India's meteorological division in predicting the future of air quality and its reputation. The accuracy of the model, as determined by R- square,is 71.69% for train data and 72.10% overall. Together with that, it is anticipated that modules like figuring out which pollutant caused the value, which age group of people is affected the most, what precautions need to be taken, and what preventative measures to lower the AQI need to be displayed on the model, would be included.

## FUTURE WORK:-

1. The model must determine the pollutant that caused the value to arise.
2. The two questions are which age groups are most impacted and what safety precautionsare to be taken.
3. The model would do well to display the preventive steps being done to reduce the AQI.

**REFERENCES**

**TEXT BOOK :-**

- Geron Aurelien. Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems, 2nd Edition, Kindle Edition, O'Reilly Media, 13 March 2017.
- Gary B. Shelly. Systems Analysis and Design, Shelly Cashman Series' textbooks, 1991.
- Oliver Theobald. Machine Learning for Absolute Beginners: A Plain English, Independently Published, 2017.

**WEBSITES:-**

- https://www.javatpoint.com/linear-regression-in-machine-learning.
- https://www.javatpoint.com/machine-learning-random-forest-algorithm.
- https://www.youtube.com/watch?v=nxFG5xdpDto.

**RESEARCH PAPERS :-**

1. K. Mahesh Babu, J. Rene Beulah. Air Quality Prediction based on Supervised Machine Learning Methods. International Journal of Innovative Technology and Exploring Engineering (IJITEE), July2019.
2. Khalid M O Nahar, Mohammad Ashraf Ottom, Fayha Alshibli and Mohammed M. Abu Shquier. AIR QUALITY INDEX USING MACHINE LEARNING. COMPUSOFT, An international journal of advanced computer technology, September-2020.
3. Ritik Sharma, Gaurav Shilimkar, Shivam Pisal. Air Quality Prediction by Machine Learning. International Journal of Scientific Research in Science and Technology, May-June-2021.
4. Avnish Bora, Laxmi Chaudhary. Technological Advancements in Air Pollution Monitoring Systems. International Journal of Engineering Research and Management(IJERM), November 2020.
5. Udit Ranjan Kalita, Heniel Kashyap, Amir Chetri and Jesif Ahmed. Centralized Air Pollution Detection and Monitoring. ADBU Journal of Electrical and Electronics Engineering (AJEEE), February 2018.
6. Yasin Akın Ayturan, Zeynep Cansu Ayturan and Hü seyin Oktay Altun. Air Pollution Modelling with Deep Learning. Int. J. of Environmental Pollution & Environmental Modelling, September 20, 2018.
7. B Mark, R Soria, S Berres, L Caro , A Mellado and N Schiappacasse. Detection of Anomalous Pollution Sensors Using Deep Learning Strategies. IOP Conf. Series: Earthand Environmental Science, June 06, 2020.
8. M Rogulski and A Badyda. Current trends in network based air quality monitoring systems. 2nd International Conference on the Sustainable Energy and Environmental Development. IOP Publishing, 2019.
9. Tomasz Cieplak, Tomasz Rymarczyk and Robert Tomaszewski. A concept of the air quality monitoring system with machine learning methods to detect data outliers. MATEC Web of Conferences 252, 2019.
10. Elias Kalapanidas and Nikolaos Avouris. Applying Machine Learning Techniques in Air Quality Prediction. International Journal of Emerging Trends & Technology in Computer Science (IJETTCS), 2018.