WWW.IJRPR.COM

**International Journal of Research Publication and Reviews**

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

# Air Quality Index Prediction Using Machine Learning

***I. Ravi***

Department of Computer Science, K. R. College of Arts and Science, Kovilpatti, India

**ABSTRACT:**

Air quality is a critical environmental parameter that directly affects public health and well-being. Monitoring and predicting the Air Quality Index(AQI)play a crucial role in assessing the level of pollution and implementing effective mitigation strategies. This study involves the Collection of historical data on pollutants $NH_3$, CO, $O_3$, etc, as well as the AQI labels that are good, satisfactory, moderate, poor, or very poor. Various regression algorithms, including Linear Regression and Lasso Regression, are employed to develop predictive models and find the best one to accurate the results. The predictive models have split the dataset as trained and validated us in the methodology to ensure their reliability and generalization capability. The performance metrics, suc has Mean Absolute Error, R Squared, and Root Mean Square Error, assess the effectiveness of the models in predicting different AQI categories. The implications of accurate AQI predictions for environmental monitoring include improved early warning systems and more effective communication of potential health risks to the public.

Keywords: Air Quality Index (AQI),regression algorithm, prediction

## 1. Introduction:

Air pollution is a significant global concern affecting public health and the environment. Monitoring and predicting the Air Quality Index(AQI) play a crucial role in assessing the level of pollution and implementing effective mitigation strategies. This study involves the Collection of historical data on pollutants $NH_3$, CO, $O_3$, etc, as well as the AQI labels that are good, satisfactory, moderate, poor, or very poor. Various regression algorithms, including Linear Regression and Lasso Regression, are employed to develop predictive models and find the best one to accurate the results. The predictive models have split the data set as trained and validated using the methodology to ensure their reliability and generalization capability. The performance metrics, such as Mean Absolute Error, RSquared, and Root Mean Square Error, assess the effectiveness of the models in predicting different AQI categories. The implications of accurate AQI predictions for environmental monitoring include improved early warning systems and more effective communication of potential health risks to the public.

## 2. Literature Review:

This paper[1] has predicted the hourly concentration values for the ambient air pollutants NO2, SO2, PM10, PM2.5, CO and O3 for the stations Naderi, Havashenasi, MohiteZist and Behdashtin Ahvaz,Iran which is the most polluted city in the world. They have also calculated and predicted Air Quality Index (AQI) and Air Quality Health Index(AQHI) for the four air quality monitoring stations in Ahvaz mentioned above. They used Artificial Neural Network (ANN) machine learning algorithm for the prediction of air pollutants concentration (hourly) and two airqualityindicesAQIandAQHIovertheAugust2009to August2010.InputtoANNalgorithmsinvolvesthefactors such as meteorological parameters, Air pollutants concentration, time and date.

This paper [2] had employed the machine algorithms to detect and forecast the PM2.5 concentration level on the basis of dataset containing atmospheric conditions in a specific city. They also predicted thePM2.5 concentration level for a particular date. First of all they classify the air as polluted or not polluted by using Logistic Regression algorithm and then Auto Regression algorithm was used to predict the future value of PM2.5 depending upon previous records.

This paper [3] have analysed the proportion of various air pollutants(NO,NO2,CO,PM10 andSO2)with respect to the time of the day and the day of the week and estimated the effect of environmental parameters as temperature, wind speed and humidity on the air pollutants mentioned above with the help of WEKA tool. The data was collected From pollution control board of Karnataka. By using Zero R algorithm in WEKA tool the study come up with the results that shows that the concentration levels of air pollutants increase during the working days and especially during the peak hours of the day and decrease during week-ends or holidays. Using Simple K-means Clustering algorithms the study shows the relationship or dependencies between the environmental factors like Temperature, wind speed and humidity and the air pollutants like NO, NO2, PM10, CO and SO2.

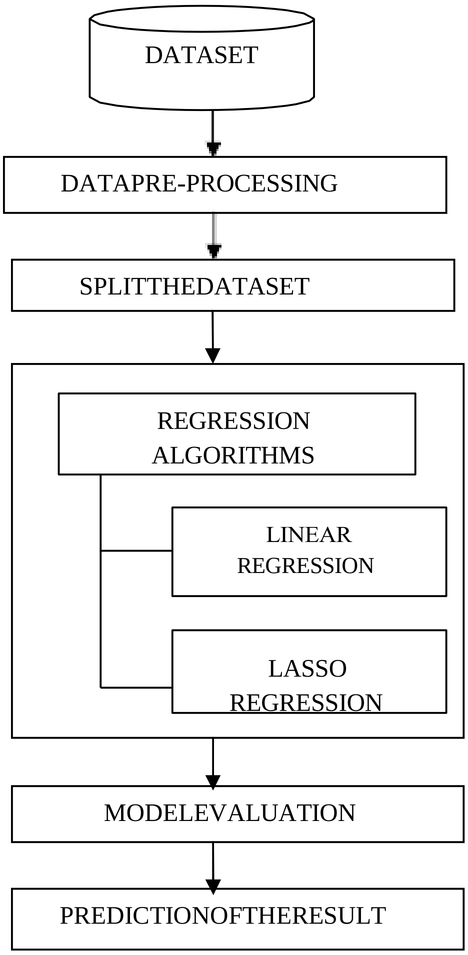
This paper [4] had used the various machine learning algorithms to predict the PM2.5 concentration. Data were collected from the official website of Environment Protection Agency (EPA) for the city Melbourne that contains PM2.5 air parameter and they have also collected the unofficial data from Air beam which is the mobile device used to measure PM2.5 value. The machine Learning Algorithms Artificial Neural Network (ANN), Linear Regression (LR) and Long Short Term Memory (LSTM) recurrent neural network were used for the PM2.5 prediction but out of these algorithms LSTM gives the best performance ad predict the high PM2.5 value with reasonable Accuracy.

This paper [5] had used the air pollutants as PM10, NO2 and O3overtheyears2014 and2015 for Kermanshah city in Iran. They used the Air Qsoftware proposed by WHO for this purpose. The number of premature deaths for cardiovascular diseases is of 188 related to PM10, 33 related to NO2 and 83 related to O3.The results of this study indicates that if there is 10µ/m3 increase in PM10, NO2 and O3 concentration level the mortality risk will increase by 1.066, 1.012 and 1.020 respectively.

This paper [6] the research employed two soft computing algorithms Artificial Neural Network (ANN) and Genetic Programming (GP) for the prediction of future concentration levels of air pollutants such as Oxides of Sulfur (SOx), Oxides of Nitrogen (NOx) and Respirable Suspended Particulate Matter (RSPM) over the year 2005 to2011forPunecityinMaharashtrawhichisatthesecond position I list of polluted cities in India. They have developed total six models (three of each algorithm ANN and GP) based on hourly average data values of pollutants concentration spanning greater than 7 years. Out of these two algorithms GP algorithms gives the better performance than ANN.

This paper [7]that has been shown that this area has no serious pollution issues related to the pollutants as Sulfur Dioxide, Oxides of Nitrogen and Suspended Particulate Matter because their annual average concentration are within the range of national standards. But the annual average concentration of the pollutant PM10 is slightly higher than the levels of national standard. Also the monthly 24-hour average concentration of PM10 in the same year were crossed the national standard level except during July to October.

## **4.** Methodolog**y:**



## 5. REGRESSIONALGORITHMS:

REGRESSION USING LINEAR REGRESSION:

Linear regression is one of the easiest and most popular Machine Learning algorithms used for predictive analysis. It shows a linear relationship between a dependent (y) and one or more independent(x) variables. It finds how the value of the dependent variable changes according to the value of the independent variable.

Certainly, here are the steps for the linear regression algorithm in a concise format:

Step1: Data Collection and Preprocessing

Collect a labeled dataset and preprocess it, including handling missing data, and creating training and testing sets.

Step2: Model Training

Use the training set to train a linear regression model.

Step 3: Evaluate the trained model's performance using the testing dataset

Step4: Model Evaluation

Calculate the performance metrics such as Mean Absolute Error (MAE), R Squared ($R^2$), and Root Mean Square Error (RMSE).

Step5: Predict and Evaluate

Use the model to make predictions based on the input features.

REGRESSION USING LASSO REGRESSION:

Lasso regression also called Least Absolute Shrinkage regression is used in machine learning algorithms. It provides greater accuracy in predictions compared to other regression models. It is used as a subset of variables to prevent over fitting and encourage feature selection.

Step1:Data Preparation

Collect and pre process the dataset, including handling missing values, encoding categorical features, and splitting the data into training and testing sets.

Step2:ModelTraining

Train the Lasso Regression model on the training data.

Step3:ModelEvaluation

Use the trained model to make predictions on the test data and assess its performance using metrics like Mean Absolute Error (MAE), R Squared ($R^2$), and Root Mean Square Error (RMSE).

Step4: Predict and Evaluate

Use the model to make predictions based on the input features.

Step5: Deployment and Use

If the Lasso Regression model meets your performance criteria, deploy it for making predictions on new, unseen data in real-world applications.

## 5. Data set Description:

The dataset used for this Air Quality Index Prediction contains a comprehensive collection of various pollutants such as PM 2.5, PM 10, NO, $NO_{2,}NO_x$, $NH_3$, CO, $SO_2$, $0_3$, Benzene, Toluene, Xylene.

Below is a detailed description of the dataset features and then type

| S.NO | Attribute Name | Type |
|---|---|---|
| 1 | City | Text |
| 2 | Date | Date |
| 3 | PM2.5 | Numeric |
| 4 | PM10 | Numeric |
| 5 | NO | Numeric |

| | | |
|---|---|---|
| 6 | $NO_2$ | Numeric |
| 7 | $NO_x$ | Numeric |
| 8 | $NH_3$ | Numeric |
| 9 | CO | Numeric |
| 10 | $SO_2$ | Numeric |
| 11 | $O_3$ | Numeric |
| 12 | Benzene | Numeric |
| 13 | Toluene | Numeric |
| 14 | Xylene | Numeric |
| 15 | AQI | Numeric |
| 16 | AQI_Bucket | Text |

The Air Quality dataset forecasting relies on several key features to provide valuable in sights into atmospheric conditions.

The features are:

The city and Date are the critical elements, serving as the name of the city and where the pollutants are present.

PM 10 and PM 2.5 are the Particular Matter are crucial pollutants affecting air quality Nitrogen Oxides (NO, $NO_2$, $NO_x$), Ammonia ($NH_3$), Carbon Monoxide (CO), Sulfur Dioxide ($SO_2$), and Ozone($O_3$) are the common air pollutants.

Benzene, Toluene, and Xylene are associated with industrial activities and can contribute to the air.

AQI (Air Quality Index) is the target variable that you want to predict. It represents the overall air quality based on the concentrations of different pollutants.

AQI_Bucket categorizes air quality into different buckets such as Good, Moderate, Satisfactory, Poor, or Very Poor.

### *5.1. Data Preprocessing:*

Identify and locate hidden values. Clean the dataset columns by removing or replacing missing or invalid entries.

## 6. Proposed Method:

Gather a comprehensive dataset containing historical information on air pollutants, including features like PM 10, PM 2.5, $No_2$, etc., and predict the labels representing the human health conditions (e.g., Good, Satisfactory, Moderate, poor, and Very Poor)

Address any missing values in the dataset through techniques such as imputation or data removal and replacing the missing values.

Convert categorical air quality index labels into numerical representations, such as encoding, to make them suitable for machine learning algorithms Divide the dataset into two parts: a training set and a testing set. The training part will comprise (70%), while the testing part will be (30%).

Experiment with various Regression algorithms, including Linear Regression and Lasso Regression.

Evaluate the algorithm's performance on the training data using appropriate evaluation metrics.

Train the chosen Regression algorithms on the training dataset.

## 7. Result Analysis:

When analyzing air quality index (AQI) prediction models, various performance metrics help assess their effectiveness. These metrics include Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R Squared ($R^2$).MAE assesses the average magnitude of prediction errors and how well the model's predictions align with actual AQI values. RMSE provides a measure of the overall accuracy of the AQI predictions, with a focus on penalizing larger errors. $R^2$ evaluates how well the model explains the variance in AQI values. A higher $R^2$ indicates a better fit. The machine learning models are validated by comparing the performance metrics. The lower the MAE, and RMSE and higher the r-squared, the machine learning model performs better.

Comparison of two algorithms:

| S.NO | MODEL | MAE | RMSE | R2 |
|---|---|---|---|---|
| 1 | Linear Regression | 0.42 | 0.57 | 0.71 |
| 2 | Lasso Regression | 0.42 | 0.61 | 0.77 |

## 8. Conclusion and Future Scope:

The primary goal of this research is to employ machine learning techniques and compare their effectiveness in solving real-world problems. The Kaggle datasets are used to assess the performance of two regression algorithms: Linear regression and Lasso regression. Lasso regression emerged as the most effective and superior performance compared to other methods. Consequently, the Lasso regression is used for making predictions.

Further, the field of air quality index prediction uses real-time data analysis using the cloud to obtain better outcomes for greater performance as the data updates for every particular interval of time. These innovations contribute to a better understanding of air quality dynamics, enabling proactive measures for mitigating environmental impact and, ultimately, benefiting society as a whole.

**References:**

[1] Heidar Malek, Armin Sorooshian, Gholamreza Goudarzi, Zeynab Baboli, Yaser Tahmasebi Birgani, Mojtaba Rahmati, “Air pollution prediction by using an artifcial neural network model”, Clean Technologies and Environmental Policy, (2019) 21:1341–1352.

[2] Aditya C R, Chandana R Deshmukh, Nayana DK, Praveen Gandhi Vidyavastu, “Detection and Prediction of Air Pollution using Machine Learning Models”, International Journal of Engineering Trends and Technology (IJETT) – volume 59 Issue 4 – May 2018

[3] Mohamed Shakir, N. Rakesh, “Investigation onAir Pollutant Data Sets using Data Mining Tool”,IEEE Xplore Part Number:CFP18OZV-ART; ISBN:978-1- 5386-1442-6.

[4] Ziyue Guan and Richard O. Sinnot, “Prediction of Air Pollution through Machine Learning on thecloud”, IEEE/ACM5th International Conference onBig Data Computing Applications and Technologies (BDCAT), 978-1-5386-5502-3/18/$31.00 ©2018IEEE DOI 10.1109/BDCAT.2018.00015.

[5] YusefOmidiKhaniabadi, GholamrezaGoudarzi, Seyed Mohammad Daryanoosh, Alessandro Borgini, AndreaTittarelli,AlessandraDeMarco,“Exposure to PM10, NO2, and O3 and impacts on humanhealth”, Environ SciPollut Res, 2016.

[6] S.TikheShruti, K.C.Khare, S.N.Londhe, “Forecasting Criteria Air Pollutants Using DataDriven Approaches: An Indian Case Study”, InternationalJournalofSoftComputing8(4),305- 312,2013,ISSN:1816-9503.

[7] R. Gunasekaran, K. Kumaraswamy, P.P. Chandrasekaran, R. Elanchezhian, “MONITORINGOFAMBIENT AIR QUALITY INSALEMCITY, TAMIL NADU”, International Journal of Current Research, ISSN: 0975-833X, Vol. 4, Issue, 03,pp.275- 280, March, 2012.