Quest Journals
Journal of Architecture and Civil Engineering
Volume 7 ~ Issue 11 (2022) pp: 01-07
ISSN(Online) : 2321-8193
www.questjournals.org

Research Paper

Prediction of Air Quality Index by Using Machine Learning

Pakki Pavan Sai[1], Raghupathi RajaSekhar[1] ,Maila Vijay Babu[1], Peddi VenkataRatnam[1]

[1]*Department of Civil Engineering ,GMRIT. Rajam ,532127.*

ABSTRACT: *In these days Air pollution is one of the main threats for developed societies. According to the WHO pollution is the main cause of deaths among children under five. Prediction of Air Quality is necessary step to be taken by government as it becoming the major concern among the health of human beings. Air Quality Index measures the quality of air. Various air pollutants causing air pollution are carbon dioxide, nitrogen dioxide, carbon monoxide. That are released from burning of natural gas, coal, and wood, industries, vehicles etc. Air pollution can cause severe disease like lungs cancer, brain disease and even lead to death. Machine learning algorithms helps in determining the air quality index. Various research is being done in this field but the results are still not accurate. The required data sets are available from the many websites like air quality monitoring sites and they are divided into training and testing. Machine learning algorithms used for prediction are Linear regression, Decision tree, Random Forest, Artificial neural network, Support vector machine. By using all these algorithms, we can predict the quality of air of a particular state or particular area.*

KEYWORDS: *Linear regression, Decision tree, Random Forest, Artificial neural network, Support vector Machine.*

R*eceived 20 Oct., 2022;* ***R****evised 01 Nov., 2022;* ***A****ccepted 03 Nov., 2022 © The author(s) 2022.*
P*ublished with open access at* ***www.questjournals.org***

I. INTRODUCTION:

Air pollution: It is a contamination of air due to presence of substance in atmosphere that are harmful to health of humans and other living beings and cause damage to climate materials.
There are various causes of air pollution.
1) out door pollution
2) indoor pollution
Not just humans are harmed by air pollution; many animals, forests, crops are as well. Therefore, a project was created to analyze the air quality using various machine learning approaches.
There are different methods in machine learning technique. It can be used for appropriate data analysis, air quality prediction and air quality determination. "AQI" Air quality index is a number that indicates how polluted the air is at given time and given place. There are many pollutants that cause air pollution . They are

Primary pollutants:
1) Carbon dioxide (CO2)
2) Sulphur oxide (SOX)
3) Nitrogen oxide (NOX)
4) Carbon monoxide (CO)
5) Chlorofluorocarbons (CFC)

•**Carbon dioxide**: carbon dioxide is playing an important role in causing air pollution .It is also named as Greenhouse gas. Global Warming a major concern caused by increase in carbon dioxide in air. CO2 is exhale by human. CO2 is also released by burning of fossil fuels.
• **Sulphur oxide** (SOX) : sulphur dioxide(SO2) released by burning coal and petroleum . It is released by various industries. When react with catalyst(NO2),results in H2SO4 causing acid rain that forms in major cause of air pollution.

• **Nitrogen Oxide** (NOx): Most commonly Nitrogen Dioxide(NO2) that is caused by thunderstorm, rise in temperature.
• **Carbon Monoxide** (CO): Carbon Monoxide is caused by burning of coal and wood. It is released by vehicles. It is odorless , colorless, toxic gas.
•**Chlorofluorocarbons**(CFC): Chlorofluorocarbons released by air conditioners, refrigerators which react with other gases and damage the ozone layer. Therefore, Ultraviolet rays reach the earth surface and thus cause harm to human beings.

Secondary pollutants:
1) Ground level ozone
2) Acid rain

Ground level Ozone: It is just above the earth surface and forms when hydrocarbon reacts with Nitrogen oxide in the sunlight presence.
Acid Rain: When Sulphur Dioxide reacts with Nitrogen Dioxide. Oxygen and water in air thus causing acid rain and fall on ground in dry or wet form.

Other pollutants:
1.PM 2.5
2.PM 10

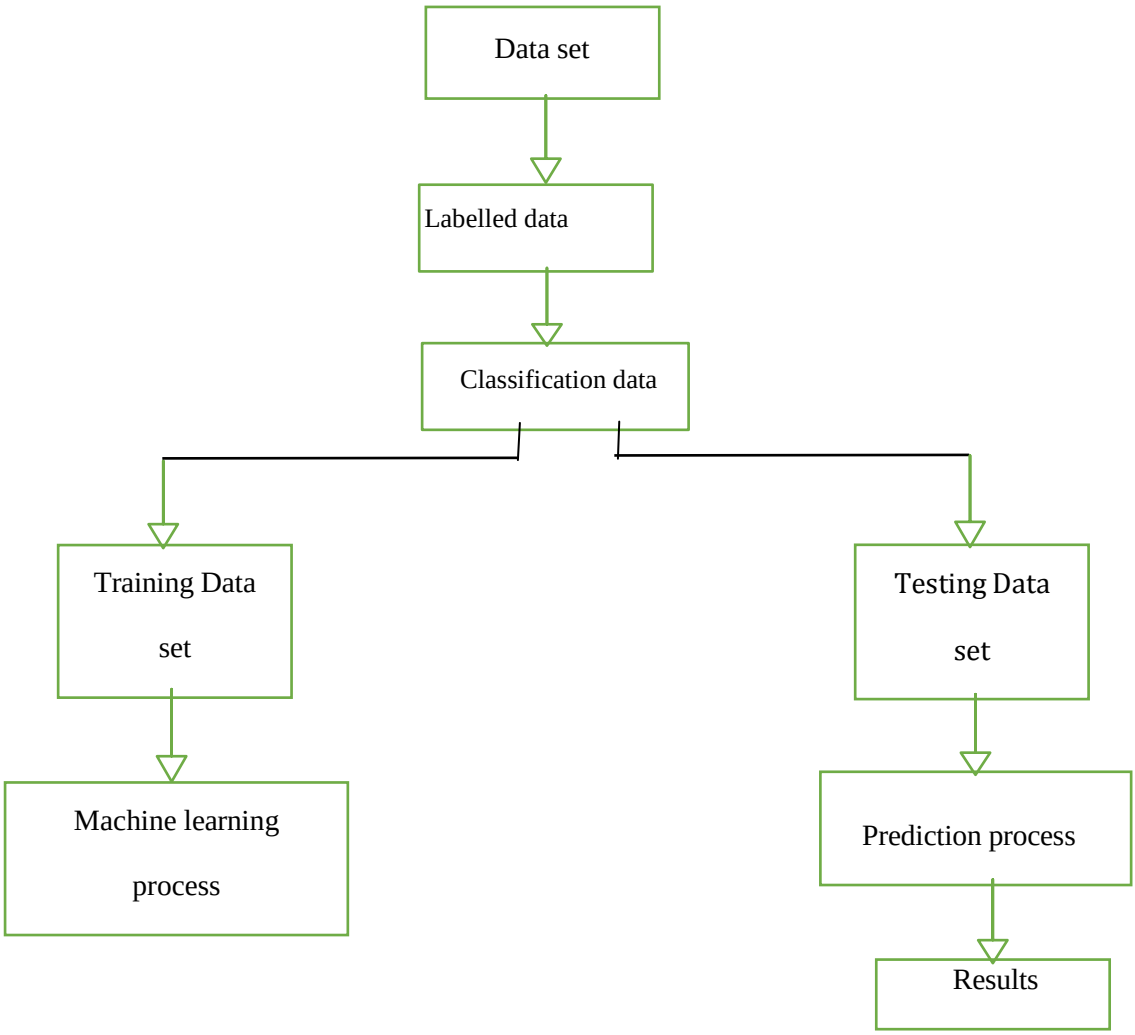
• **PM2.5:** It have a diameter of 0.0025mm, Fine particulate matter is an air pollutant that is concern for people's health when levels in air are high. PM2.5 are tiny particles in the air that reduce visibility.
• **PM10**: These are small particles found in dust and smoke. They have a diameter of 10micrometres (0.01mm) or smaller. PM10 particles are a common air pollutant.

Machine Learning prediction methods:
Machine learning involves computational methods which learn from complex data to build various models for prediction. The study attempts to build forecasting models capable of efficient pattern recognition. In this section the underlying principle of machine learning methods as their procedure will be discussed respectively. Which help us to find out the AQI.

II. Methodology

There are two primary phases in this system.
1. Training phase: The system is trained using the data in the data set and fits a model based on algorithm chosen accordingly.
2. Testing phase: The system is provided with the inputs and it is tested for its working. From the result the accuracy will be checked

Basic Algorithm for Machine Learning Methods



III. Prediction Methods:

Random forest:

The prominent machine learning method, random forest. It is a supervised ensembled algorithm (which makes data to store in a group). The method combines multiple decision trees to form a forest and the bagging concept which adds randomness to the building model. The random selection of features is used to create training data subset for each decision tree. The selection of features split individual decision tree into different parts. At each decision node in every tree, the variable from number of features is considered for best split. If the target attribute is defined, random forest will choose most frequent as its prediction. suppose the input is in numerical form, the average of all prediction will be chosen. For prediction, each test data point is passed through every decision tree in the forest. Then the trees give a vote for the best outcome from that the prediction will get the best result from the majority vote among the models. The method random forest can overcome the prediction variance and the prediction average will approximate the true value. The below figure shows the classification of random forest contains m number of trees.

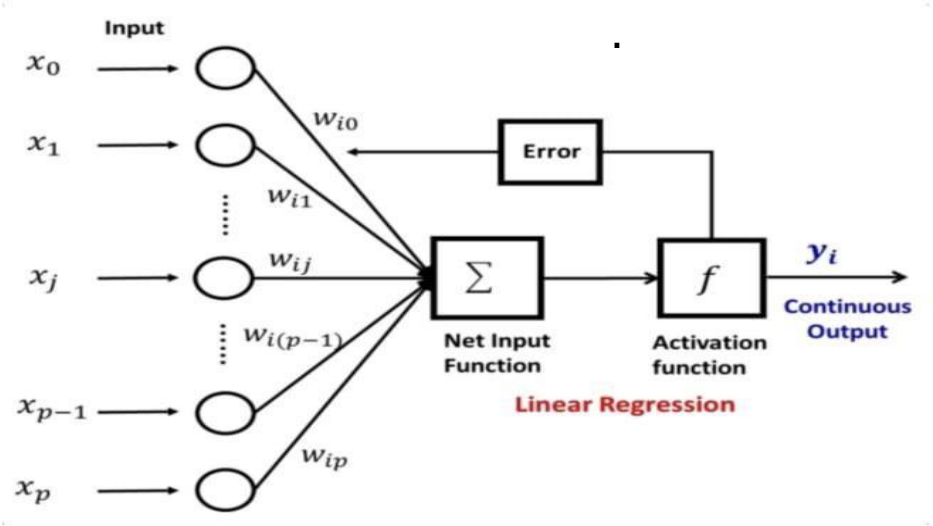
Artificial Neural Network:

It is one of the machine learning methods used to predict Air Quality Index and many other. Artificial neural network is an "Universal approximator". The method stimulates human brain in logical information. In this we used multi-layer perception (MLP) consists of processing elements called neurons. An MLP typically consists of several layers of nodes with the first layer is the input layer where it can receive external information ,and the last layer is the output layer where it can obtained result. The out layer is separate by one or more hidden layers called intermediate layers. The amount of input data is fed into Neural network where all the neurons trained and the network is used to get better output. Each neuron has specification (or)certain storage.

The neurons transmit the data from one layer to another layer. In the Neural network system, the number of hidden layers is complex by using this we can predict the quality of air using Artificial neural Network.

Linear regression:

It is used to predict the real values using continuous variables. Linear regression is probably the method where the most of the academicians started their first machine learning experience. The main principle behind this method is fitting one or more independent variables with the dependent variables into a line in n dimensions. n denotes the number of variables within the data set. The line is created which has capable of minimizing the errors when try to fit in the instantaneous line. The method linear regression is capable to optimize the parameters in the model. The optimization works on partial deriving the loss function and the parameters will be updated by subtracting the previous differentiate value. The learning rate can be done by the simplest way , which follows the rule of thumb (trail &error). Regularization is one of the parameter that has capacity to modify the model. Two types of regularization in linear regression are lasso and ridge regression .where lasso used to eliminate less important feature by letting the feature coefficient be zero and ridge regulation will not try to eliminate a feature but instead of that it try to shrink the magnitude coefficient.

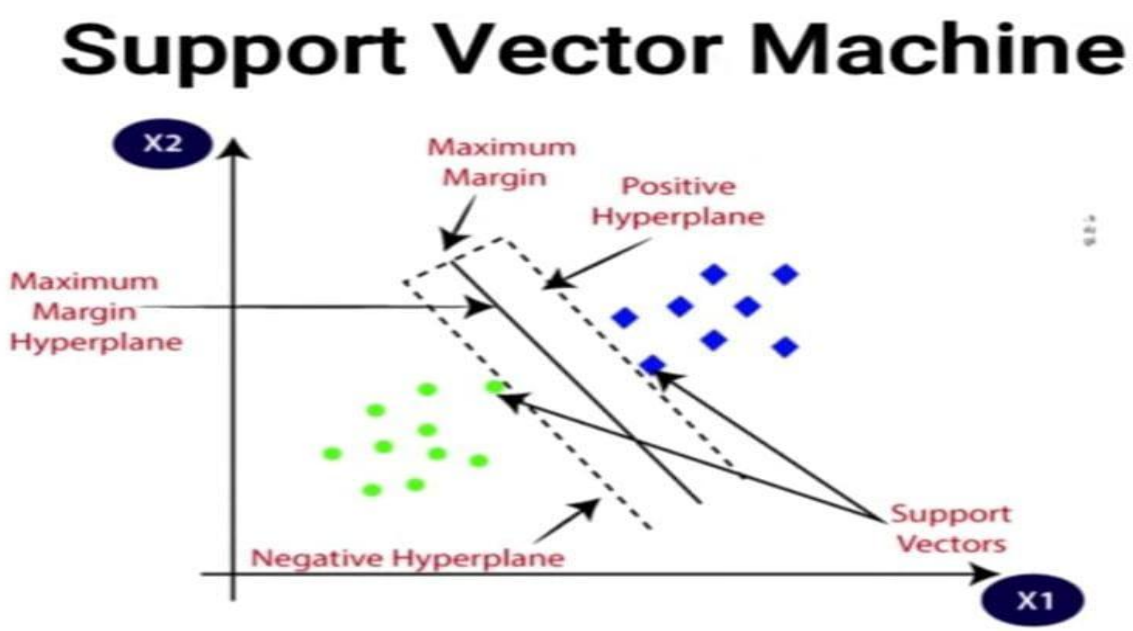


Support vector machine:

Supported vector machine is a supervised learning. In this method we will classify the regressions Hyperplane acts as a boundary between distant datapoints. There are two types of SVM .Data points that lie at the edge of an area closest to the hyperplanes are considered as support vectors .Hyperplanes will determine the number of classes incurred in the data set and the output of unseen data will be predicted .

For regression problem an approximation of hyperplane to a nonlinear function is constructed at the maximal margin with linear regression SVR uses a penalty concept introduced by parameter C (cost factor) for output variables outside the boundaries.

Support vectors represents the data points located a near boundary lines. If the ξ Moves further from the hyperplane the number of support vectors decreases otherwise the number of support vectors increases as the ξ approximately towards the hyperbola.



CLASSIFICATION OF AIR QUALITY INDEX:

SNO	AQI	REMARK	HEALTH IMPACT
1	0-50	Good	Minimal impact
2	51-100	Satisfactory	Minor breathing discomfort
3	101-200	Moderate	Breathing discomfort to the people with lungs, asthma and heart diseases
4	201-300	Poor	Breathing discomfort to most people on prolonged Exposure
5	301-400	Very poor	Respiratory illness on prolonged exposure
6	401-500	Severe	Affects healthy people and seriously impacts those with existing diseases

Implementation methodology:
The methodology in this study consists of two procedures, they are Data collection and data pre-processing.

AQI Calculation:
• AQI is calculated based on averages of all pollutant concentration measured in a full hour , a full 8 hour, or a full day. To calculate an hourly Air Quality Index, we average at least 90 measured data points of pollution concentration from a full hour.
• 24 Hourly average concentration value (8-hourly in case of CO and O3)

Data collection:
In the part of data collection firstly we have to check where the data will be available in an open source. Later we have to collect all the data required for prediction process like pollutants (PM 2.5, PM 10, SO2, CO2). With respect to the survey we should decide the place where to predict the quality of air, so that we can help our future generation to maintain a good environment.

Data pre-processing:
In the part of data pre-processing using machine learning algorithm, we have to entered all the data collected in the previous step into the selected method. After entering the data, we can find out the Air quality index by solving and comparing with some parameters, like RMSE, MAE.

Performance evaluation:
The most used parameters are RMSE (root mean square error) and MAE (mean average error), calculated based on the difference between the prediction result and the true value. Here the value of RMSE is used to explain the

strength of relation ship between predictive models. These parameters provide an idea for comparative analysis of different parameters for each model for different methods. In this process each model is re-built 20 times using different subsets of training and testing

Algorithm	**Accuracy**
Random forest	99.19%
Artificial neural network	95.01%
Support vector machine	70.56%
Logistic regression	98.38%

IV. Conclusion:

By using the above methods and different algorithms we can predict the quality of air.
Which helps the future generation to take care of environment and we can also suggest GOI to take the safety measures for future generation. The above-mentioned methods will give the accurate prediction of air. From the entered data of particular place, we got the value of parameters required. Random forest algorithm is most preferable method which gives good prediction values. And support vector machine gives the worst results among the all methods mentioned above.

REFERENCES:

[1]. Jiang, Ningbo, and Matthew L. Riley." Exploring the utility of the random forest method for forecasting ozone pollution in SYDNEY." Journal of Environment Protection and Sustainable Development 1.5 (2015): 245-254.
[2]. Svetnik, Vladimir, et al. "Random forest: a classification and regression tool for compound classification and QSAR modelling." Journal of chemical information and computer sciences 43.6 (2003): 1947-1958.
[3]. Biau, GA˜ Srard. "Analysis of a random forest model." ˇJournal of Machine Learning Research 13. Apr (2012): 1063- 1095.
[4]. Gokhale Sharad and Namita Raokhande ," Performance evaluation of air quality models for predicting PM10 and PM2.5 concentrations at urban traffic intersection.Science of the total environment 394.1(2008):9-24.
[5]. Xia Xi, Zhao Wei and Rui Xiaoguan , " A comprehensive evaluation of air pollution prediction improvementby a machine learning method", IEEE International Conference on Service Operations And Logistics, And Informatics,2019. Dan wei air pollution level in a specify city(2014).
[6]. Sachit Mahajan, Ling-Jyh Chen, Tzu-Chieh Tsai : An Empirical Study of PM2.5 Forecasting Using neural network.
[7]. Ayele, TemeseganWalelign, and RutvikMehta."Air pollution monitoring and prediction using IoT." In 2018 Second International Conference onInventive Communication and Computational Technologies (ICICCT), pp. 1741-1745. IEEE,2018.
[8]. Carbajal-Hernández, JoséJuan"Assessment G.Lind, "Air quality prediction using optimal neural networks with stochastic variables", and prediction of air quality using fuzzy logic Atmospheric Environment 79(2013): 822-830.and autoregressive models."
[9]. Atmospheric [12] Challa Venkara Srinivas et al ," Data Environment 60 (2012):37-50.
[10]. Yu Jiao, Zhifeng Wang, Yang Zhang, Prediction of air quality index based on LSTM, IEEE,2019.
[11]. Veljanovska, K.; Dimoski, A. Air Quality Index Prediction Using Simple Machine Learning Algorithms. Int. J.Emerg. Trends Technol. Comput. Sci. 2018, 7, 25–30.
[12]. [Afshar-Mohajer et al.2018] N.zuidema,C.Sousan,S., Hallett,L.,Tatum,M.,Rule ,A.M.,Thomas,G.,Peters ,T., and Koehler,k.(2018). Evaluation of low cost electro chemical sensors for environmental monitoring of ozone,nitrogen dioxide and carbon monoxide. Journal of occupation and environmental hygiene.
[13]. Gokhale sharad and Namita Raokhande, "Performance evaluation of air quality models for predicting PM10 and PM2.5 concentrations at urban traffic intersection during winter period", Science of the total environment 394.1(2008):9 24.
[14]. Bhanarkar, A. D., et al, "Assessment of "Atmospheric Environment 39.40(2005):7745-India." Atmospheric Environment 39.40 (2005):7745-7760.
[15]. Singh Kunwar P., Shikha Gupta and Premanjali Rai, " Identifying pollution sources and prediction urban air quality using ensemble learning methods", Atmospheric environment80 (2013): 426-437.
[16]. Wang Jun, and Sundar A. Christopher, "Intercomparison between satellite derived aerosol optical thickness and PM2. 5 Mass: Impliances for air quality studies", Geophysical research letters30.21(2003).
[17]. Sharma M E A McBean and U.Ghosh, "Prediction of atmospheric sulphate deposition at sensitive receptors in northernIndia",Atmospheric Environment 29.16(1995) 2162.
[18]. Wang Z et al , " A nested air quality prediction modelling system for urban and regional scales : Application for high high-ozone episode in Taiwan " Water, Air and Soil Pollution130.1-4(2001):391-396.
[19]. Ghorani-Azam, A.; Riahi-Zanjani, B.; Balali-Mood, M. Effects of Air Pollution on Human Health and Practical Measures for Prevention in Iran. J. Res. Med. Sci. 2016, 21, 1–12.
[20]. Veljanovska, K.; Dimoski, A. Air Quality Index Prediction Using Simple Machine Learning Algorithms. Int. J. Emerg. Trends Technol. Comput. Sci. 2018, 7, 25–30.
[21]. Shrestha, D.L.; Solomatine, D.P. Machine Learning Approaches for Estimation of Prediction Interval for the Model Output. Neural Netw. 2006, 19, 225–235.
[22]. Yamamoto SS, Phalkey R, Malik AA (2014). A systematic review of air pollution as arisk factor for cardiovascular disease in South Asia: Limited evidence from India and Pakistan.
[23]. Cohen AJ, Ross Anderson H, Ostro B, Pandey KD, Krzyzanowski M, Künzli N, Gutschmidt K, Pope A, Romieu I, Samet JM, Smith K (2005) The Global Burden of Disease Due to Outdoor Air Pollution. J. Toxicol. Environ. Health 68(13–14), 1301– 1307.
[24]. Dominici F, Peng RD, Bell ML, Pham L, McDermott A, Zeger SL, Samet JM (2006). Fine Particulate Air Pollution and Hospital Admission for Cardiovascular and Respiratory Diseases. JAMA 295(10):1127.
[25]. Smith KR (2000). National burden of disease in India from indoor air pollution. P Natl Acad Sci USA 97(24):13286–13293.

[26]. Ali M, Athar M (2007) Air pollution due to traffic, air quality monitoring along three sections of National Highway N-5, Pakistan. Environ. Monit. Assess136(1-3):219–226.
[27]. Tasic M, Rajsic S, Novakovic V, Mijic Z (2006) Atmospheric aerosols and their influence on air quality in urban areas. Facta Universitatis - Series: Physics, Chem Technol 4(1):83–91.
[28]. Sharma, R., Kumar, R., Sharma, D.K. et al. Inferring air pollution from air quality index by different geographical areas: case study in India. Air Qual Atmos Health 12, 1347– 1357 (2019).
[29]. Mir KA, Purohit P, Goldstein GA, Balasubramanian R (2016). Analysis of baseline and alternative air quality scenarios for Pakistan: an integrated approach. Environ. Sci. Pollut. Res 23(21):21780–21793.
[30]. Tabinda AB, Ali H, Yasar A, Rasheed R, Mahmood A, Iqbal A (2019) Comparative Assessment of Ambient Air Quality of Major Cities of Pakistan. MAPAN 35(1):25–32.