AIR POLLUTION PREDICATION USING MACHINE LEARNING

Kalash Agarwal[1], Yatender Singh[2], Jasmendra Singh, Abhishek Goyal[4]

[1,2,3]*Student, Computer Science & Engineering Department, ABES Engineering College, Ghaziabad, U.P., India*
[4]*Professor, Computer Science & Engineering Department, ABES Engineering College, Ghaziabad, U.P., India*

---***---

Abstract - *In the populated and developing countries, governments consider the regulation of air as a major task. Monitoring air quality is a necessary exercise in the meteorological and movement factors, stubble burning and open construction practice these factors contribute a lot in air pollution. So forecast air quality index using a machine learning model to predict air quality index for NCR(national capital region). The values of major pollutants like SO2, PM2.5, CO, PM10, NO2, and O3.in recent years machine learning in most emerging technology for predicting on historical data with 99.99% of accuracy. we implemented different classification and regression techniques like Linear Regression, multiple linear regression, KNN, Random Forest Regression, Decision Tree Regression, Support Vector Regression, Artificial Neural Networks. To make more accurate our prediction use Mean square error, mean absolute error and R square. To prognosticating air quality index of NCR(national capital region) in different aspects of like stubble farming, Motor vehicle emission, and open construction practice which result in the air quality of NCR.*

Key Words: Machine learning, air pollution prediction, linear regression, artificial neural network, KNN, SO2, PM2.5.

1. INTRODUCTION

Introduction of contaminants into the natural environment that causes adverse change. Pollution can take the form of energy, such as noise, heat or light. Pollutant particle, the components of pollution, can be either foreign substances/energies or naturally occurring contaminants. Pollution is often classified as a single source or more than one source. The main four types of pollution are Water pollution, soil pollution, noise pollution and air pollution. We here elaborate detail about air pollution. A factor which induces air pollution is stubble farming, motor vehicle emission, topological factor, and open construction work. The ordinance of environmental contamination has drawn public examination. NCR (National Capital Region) one of the most contaminated territories in the world. Different researches have been carried to evaluate the air contamination trends and its dangerous consequences like one current research have guided that the concentration of pollutants in NCR is extremely higher than the allowable limits [1]. This has resulted in a lessening in the living anticipation of its citizens by six years. While another [2] published the results of air vehicular contamination on human fitness. hence, we enhancing the air quality forecasting is one of the best objectives for civilization. Sulphur dioxide, PM2.5 and NO are major pollutants found in the air. Sulphur Dioxide is a gas. present in air. This combines easily with different substances to form harmful substances like Sulphur acid, sulfurous acid etc. Sulfur dioxide affects social fitness when it is inhaled in. It burns the nose, throat, and airways to provoke coughing, wheezing, the brevity of breath, or a tense feeling around the chest. The concentration of Sulphur dioxide in the environment can affect the habitat appropriateness [3].PM2.5 is also known as fine particulate matter (2.5 micrometers is one 400th of a millimeter). Fine particulate matter (PM2.5) is significant among the pollutant index because it is a big concern to people's health when its level in the air is relatively high [4]. So, we categorized according to Air quality index table.

2. RELATED PREVIOUS WORK

Numerous Many air quality forecast models exist to assess and prognosticate pollutant concentrations in metropolitan cities. Traditionally analytical models and statistical models include synthetic variation models and atmospheric dispersal models were applied for prognostication. Recently machine learning techniques have appeared as the outstanding methods used in air quality prognostication models.

MACHINE LEARNING MODELS

As we know that artificial intelligence makes an impact on our day to day life, artificial intelligence-based algorithms are being widely used for prediction purposes, especially for air quality forecasting. A machine learning approach takes into account multiple parameters for prediction, unlike a purely statistical model. Artificial Neural Networks (ANNs) have emerged to be the several broadly accepted methods for the prognostication of air quality [5]. Different investigations have presented the use of composite models or different models based on regression algorithm for prediction. Artificial intelligence algorithms such as fuzzy logic, generative algorithm, Principal component analysis (PCA) along with ANNs have been applied to create such models like Adaptive Neuro Fuzzy Interface System (ANFIS) model [5] etc. Another machine learning models that have been recognized add Support Vector Machine (SVM) situated model [1], PCA-SVM [18] and several also. A modified Lasso and Ridge regression technique mode0l [19] where K-nearest neighbour algorithm has also been implemented to determine concentrations of PM2.5, SO2 and PM10. Another study conducted in Quito, Ecuador [6], worked on six meteorological constituents for predicting AQI concentrations. K. Hu et al. [5], planned a machine training model Haze Est for prognosticating the air index. Here, first, the method was evaluated using seven distinctive regression technique and finally, SVR (Support vector regressor) was chosen as the ultimate prognostication model. The main goal is to prognosticate an air contamination level in an urban area with the ground data set.[4]

This method has used the Linear regression and Support vector regression for the forecast of the contamination of the next month, the next day and any date of future. The method improves to prognosticate any date contamination details within one period based on independent parameters and examining pollution parts and determine future pollution. Time Series Analysis was also used for the identification of future data points which have seasonality and trends in air pollution prediction.[3]

This designed method performs two significant tasks (i). Identifies the levels of pollutants (S02, PM2.5, CO, benzene) based upon provided meteorological values. (ii) Prognosticates the level of pollutants for a special date. Logistic regression is used to identify a data sample is either contaminated or not contaminated. Autoregression is used to prognosticate projected values of pollutants based upon the early pollutants' interpretations. The prime aim is to prognosticate the air pollution level within a particular area with the raw data set.[4]

3. PROPOSED METHODOLOGY

3.1 DATA SOURCE

3.1.1 To prognosticate the air quality of The NCR area, we want the pollutant concentration of all the elements available in the air. Which will be available in the **cpcb.nic.in** the website, which holds all the data that contaminates the area every year. We use data from several stations which measures many elements present in the atmosphere. Data is taken from 10 different stations in NCR. These data are stored in the form of a table which consists of a total of 3469 rows and having 8 columns in each row. The AQI formulae will be applied in order to calculate the AQI by using the various regression algorithm for a particular year.

3.1.2 CALCULATING AQI

An air quality index (AQI) is used by government agencies to communicate to the public how polluted the air currently is or how polluted it is forecast to become. Public health risks increase as the AQI rises. Different countries have their own air quality indices, corresponding to different national air quality standards. We here focus on formula which use to calculate AQI in India

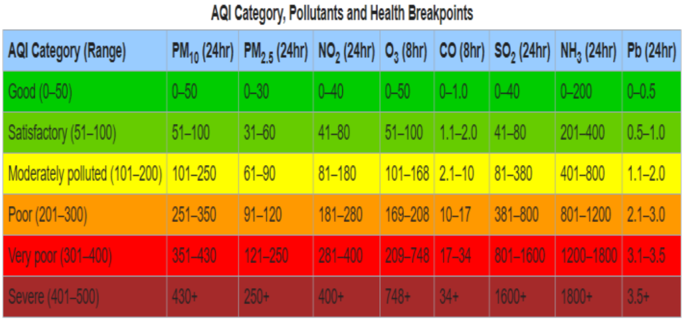
3.1.3 COMPUTING AQI

The air quality index is a piecewise linear function of the pollutant concentration. At the boundary between AQI categories, there is a discontinuous jump of one AQI unit. To convert from concentration to AQI this equation is used:

$$I = \frac{I_{high} - I_{low}}{C_{high} - C_{low}}(C - C_{low}) + I_{low}$$

Where:

C_{low}= the concentration breakpoint that is ≤ C,
C_{high}= the concentration breakpoint that is ≥ C,
I_{low}= the index breakpoint corresponding to C_{low},
I_{high} = the index breakpoint corresponding to C_{high}.



AQI Category, Pollutants and Health Breakpoints

AQI Category (Range)	PM_{10} (24hr)	$PM_{2.5}$ (24hr)	NO_2 (24hr)	O_3 (8hr)	CO (8hr)	SO_2 (24hr)	NH_3 (24hr)	Pb (24hr)
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	0.5–1.0
Moderately polluted (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	10–17	381–800	801–1200	2.1–3.0
Very poor (301–400)	351–430	121–250	281–400	209–748	17–34	801–1600	1200–1800	3.1–3.5
Severe (401–500)	430+	250+	400+	748+	34+	1600+	1800+	3.5+

Table -1: EPA table

3.2 DATA PREPROCESSING

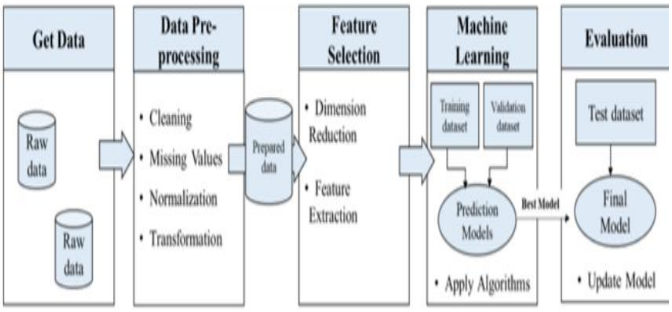
3.2.1 DATA REFINEMENT- The data to be analyzed were cleaned by discarding the instances having missing values in input parameters. The missing values in case of the target object, i.e., the pollutants is estimated using an imputer function to perform the interpolation. The strategy used here for estimation is the mean value.

3.3 FEATURE SELECTION

Feature selection is the method of choosing a subset from primary features that include important information to prognosticating output data. In the case of unnecessary data, feature extraction implies used. Feature extraction includes the choice of best input parameters of the chosen input dataset. The unified dataset hence gathered is used for further study. The maximum amount of inputs available for review is seven, hence all the inputs are selected for the computations.

3.4 TRAINING THE MODEL

Data Splitting was done as 80% for training and 20% for testing.

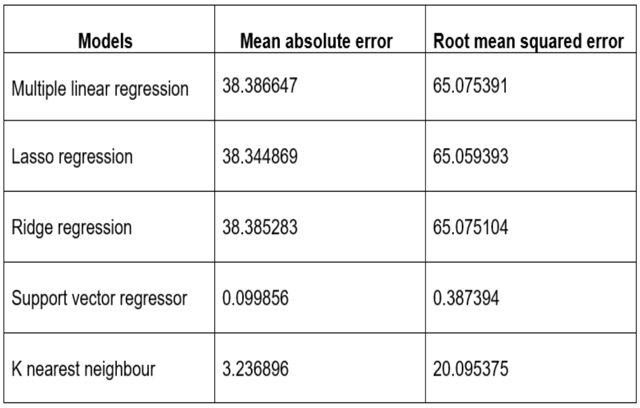


3.5 IMPLEMETATION OF DIFFERENT TYPES OF REGRESSION MODEL

In Machine learning lot of model for predicting outcomes. Some for regression and some classification. As we know that AQI distinct value so we need to train model on regression algorithm. We decided to implement different type of algorithm like support vector regression, multiple linear regression, Lasso-Ridge regression and k-nearest neighbor and etc.

3.6 EVALUATION

After implementing different types of algorithm, we need to evaluate which algorithm is best for predict AQI of NCR. Table no 2 show root mean and r squared values of algorithm which mention in above paragraph.



Models	Mean absolute error	Root mean squared error
Multiple linear regression	38.386647	65.075391
Lasso regression	38.344869	65.059393
Ridge regression	38.385283	65.075104
Support vector regressor	0.099856	0.387394
K nearest neighbour	3.236896	20.095375

Table -2: Modal evaluation

4. RESULTS AND DISCUSSION

The machine learning model to predict air pollution was one the most important objective of the research paper. We here predict air pollution on particular data i.e. 1 January 2020.We take data of previous years the data from www.cpcb.nic.in. By using linear regression, Lasso-Ridge regression, KNN and support vector machine. We actually compare actual value and predicted values.

After evaluation of different type of regression model. The best fit model for predict AQI is K-nearest neighbor model which having accuracy of 97.5 percentage. this model is best fit. As shown fig 3 comparing actual values with predicted values

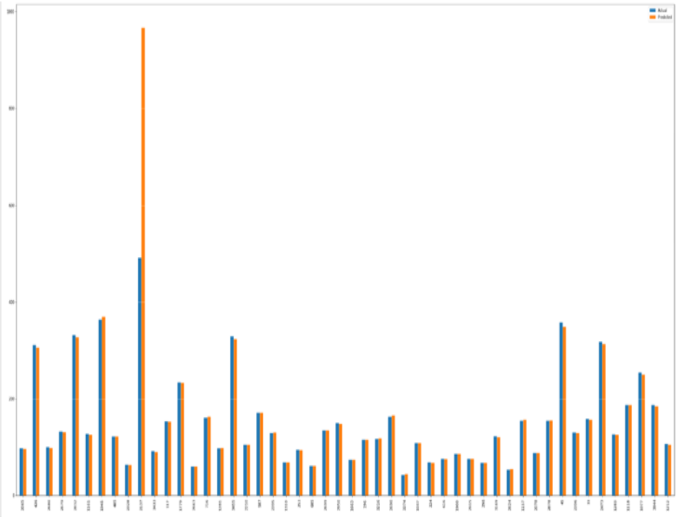
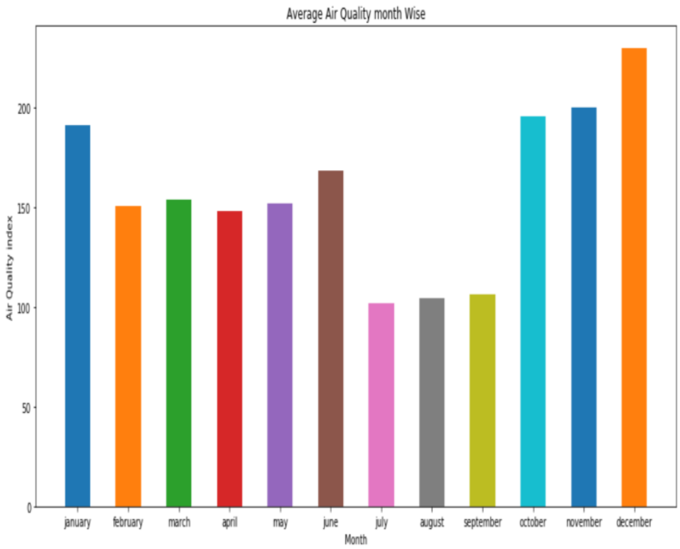


Fig -1: Comparison between actual and predicted values

Here, we show that the almost all values is equal some particular values show anomalous behavior. If we see average of AQI in month wise Fig 4 we easily say that AQI not only depends on concentration of particle it also show that AQI also depends on temperature, humidity, etc. It conclude that we need more data and Add more column in dataset of other factor like temperature and etc.



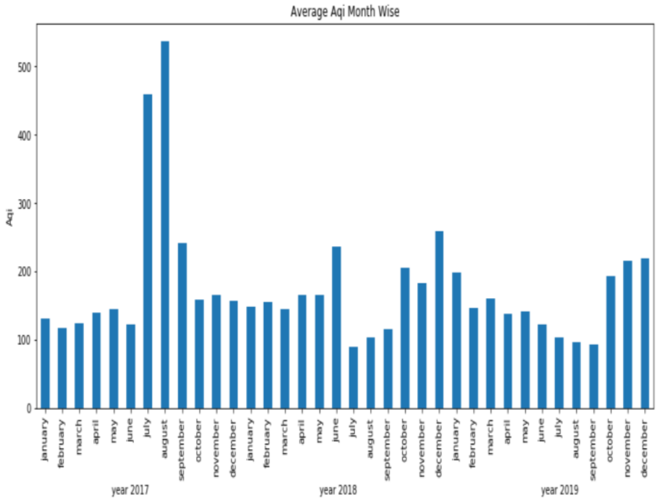


Fig -2: Average of AQI according month

We clearly see that, if we particularly show average July and august month in 2017 it is far more differ than other years same month.

5. FUTURE SCOPE

In this paper their have lots of scope in future because data was limited at this time. If we need good predicting model, we need two merge AQI prediction with Temperature prediction which gave us far good result than this model with more data this model best fit predicting in future. We need further more tiny constraint for future predicting AQI Identify the predictor(s) for which the variance is not properly captured (reason for heteroscedasticity). This will solve the problem for normality as well. Search for other avenues to look for quality controlled data. Apply models to a greater number of stations to increase the training input. We have limited data for prediction. We also train our model for next year data. Solve the problem of auto-correlations.

6. REFERENCES

1. Chavi Srivastava, Amit Prakash Singh and Shyamli Singh “Estimation of Air Pollution in Delhi Using Machine Learning Techniques “.https://www.researchgate.net/publication/332430367.”

2. P. Aggarwal and S. Jain, "Impact of air pollutants from surface transport sources on human health: A modelling and epidemiological approach", Environ. Int., vol. 83, pp. 146-157, 2015.

3. Pooja Bhalgat, Sejal Pitale and Sachin Bhoite ”Air Quality Prediction using Machine Learning Algorithms” International Journal of Computer Applications Technology and Research Volume 8–Issue 09, 367-370, 2019, ISSN:-2319–8656

4. Pandey, Gaurav, Bin Zhang, and Li Jian. "Predicting sub-micron air pollution indicators: a machine learning approach." ; Environmental Science: Processes & amp; Impacts 15.5 (2013): 996-1005.

5. M. Baawain, "Systematic Approach for the Prediction of GroundLevel Air Pollution (around an Industrial Port) Using an Artificial Neural Network", Aerosol Air Qual. Res., 2014.

6. Shyamli Singh, Amit Prakash Singh “Estimation of Air Pollution in Delhi Using Machine Learning Techniques “2018 International Conference on Computing, Power and Communication Technologies https://www.researchgate.net/publication/332430367

7. Nandini Bhalla,Janee O’Boyle “Who is Responsible for Delhi Air Pollution?Indian Newspapers’ Framing of Causes and Solutions “International Journal of Communication 12(2018),41-64”

8. Sharma AK,Baliyan P “Air pollution and public health:the challenges for Delhi,India”Reviews on Environmental Health,01 Mar 2018,33(1):77-86

9. SA Rizwan, Baridalyne Nonengkynrith, Sanjeev Kumar Gupta “Center for Community Medicine,All India Institute of Medical Sciences,New Delhi,India

10. Ishita Jalan and Hem H.Dholakia “What is Polluting Delhi’s Air? Understanding Uncertainties in Emissions Inventories. Issue Brief March 2019 www.ceew.in

11. C.B. Tripathi ,Prashant Baredar and Lata Tripathi “Air pollution in Delhi: biomass energy and suitable environment policies are sustainable pathways for health safety” ,Current Science Vol. 117,No.7,10 October 2019

12. Sachin Kumar “Air Pollution and Climate Change : Case Study National Capital Territory of Delhi.