International Journal of Scientific Research in Science and Technology
Print ISSN: 2395-6011 | Online ISSN: 2395-602X (www.ijsrst.com)
doi : https://doi.org/10.32628/IJSRST218396

Air Quality Prediction by Machine Learning

Ritik Sharma, Gaurav Shilimkar, Shivam Pisal
Vishwakarma Institute of Technology, Pune, Maharashtra, India

Article Info
Volume 8, Issue 3
Page Number : 486-492

Publication Issue
May-June-2021

Article History
Accepted : 25 May 2021
Published : 31 May 2021

ABSTRACT

The air quality observing framework estimates different air toxins in different areas to keep up great air quality. It is the consuming issue in the current situation. Air is defiled by the appearance of risky gases into the environment from the enterprises, vehicular outflows, and so forth These days, air contamination has arrived at basic levels and the air contamination level in many significant urban areas has crossed the air quality list esteem as set by the public authority. It significantly affects the soundness of the human. With the headway in innovation of ML, it is currently conceivable to anticipate the poisons dependent on the past information. In this paper we are presenting a gadget that can proceed with that can take present poisons and with the assistance of past toxins, we are running a calculation dependent on the ML to anticipate the future information of contaminations. The detected information is saved inside the Excel sheet for additional assessment. These sensors are utilized on the Arduino Uno stage to gather the contamination information.

Keywords : Machine Learning, Internet of Things, AQI, Air Pollution

I. INTRODUCTION

Air contamination observing has acquired consideration these days as it significantly affects the wellbeing of people just as on the biological equilibrium. Other than because of the impacts of harmful emanations on the climate, wellbeing, work usefulness and effectiveness of energy are additionally influenced by the air contamination. Since air contamination has caused numerous perilous consequences for people it ought to be checked persistently with the goal that it tends to be controlled adequately. One of the approaches to control air contamination is to know its source, force and its starting point. Typically, it is checked by the individual express government's current circumstance service. They keep the string of the toxin gases in the individual regions. The information introduced by the WHO is cautioning about the contamination's levels in the country. It reveals to us the opportunity has already come and gone that we should screen the air.

Air tracking manner to measure ambient ranges of air pollutants inside the air. Monitoring has become a major job as air pollution has been increasing day by day. Continuous monitoring of air pollution at a place

gives us the levels of pollution in that area. From the information obtained by the device gives us information about the source and intensity of the pollutants in that area. Using that information, we can take measures or make efforts to reduce the pollution level so that we can breathe in a good quality of air. Air pollution not only affects the ecological balance but also the health of humans. As the levels of gases increases in the air, those gases show a major impact on the human body and lead to hazardous effects. Air pollution also affects the seasonal rainfall too due to an increase of pollutants in the air. The rainfall is also affected. Hence, continuous monitoring of the air is necessary.

The major cases of air pollution are Ozone (O3), Nitrogen dioxide (NO2), Carbon Monoxide (CO), Sulphuric oxide (SO2), Particular matter (PM). These gases are cannot been seen or noticed which are produced from burning of fossil fuels, wood burnings, industrial boilers and from the explosion of volcano. They may cause the affects in humans and are the main reason for causing cancer, birth defects and breathing related problems.

Air Quality Index- Nowadays pollution levels are increasing due to the PM2.5 gases which affect the heart functionalities, lung cancer and other respiratory and breathing problems. The long-term damage to the liver, kidney, brain, nerve and other organs in the human body system is affected by air pollution. The AQI is a linear feature of the pollutant concentration. The boundaries between AQI there is discontinuous jump between AQI categories unit to other. To calculate the AQI from the concentration the below equation is used.

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

where:

I = the (Air Quality) index,
C = the pollutant concentration,
C_{low} = the concentration breakpoint that is $\leq C$,
C_{high} = the concentration breakpoint that is $\geq C$,
I_{low} = the index breakpoint corresponding to C_{low},
I_{high} = the index breakpoint corresponding to C_{high}.

Environmental protection agency breakpoint table[1]

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)
Good (0–50)	0–50	0–30	0–40	0–50	0–1.0	0–40	0–200
Satisfactory (51–100)	51–100	31–60	41–80	51–100	1.1–2.0	41–80	201–400
Moderately polluted (101–200)	101–250	61–90	81–180	101–168	2.1–10	81–380	401–800
Poor (201–300)	251–350	91–120	181–280	169–208	10–17	381–800	801–1200
Very poor (301–400)	351–430	121–250	281–400	209–748	17–34	801–1600	1200–1800
Severe (401–500)	430+	250+	400+	748+	34+	1600+	1800+

II. LITERATURE SURVEY

A brilliant air quality observing framework is proposed in [2] which detects the toxin gases at a specific region and transfer the information into the worker with the goal that everyone can know about the air contamination. The gathered information transferred to the worker anytime in the particular site. This strategy recognizes engines causing toxins and measures different kinds of poisons, and its level in air. The deliberate data is shared to vehicle proprietor and specialists of the site guests control to keep toxins made through the Air quality observing with occasion based detecting is introduced in [3]. In this work, creators demonstrated that the procedure saves half of the sensor energy utilization contrasted with customary intermittent detecting techniques dependent on the contextual investigation in the city of Spain.

The air quality qualities are anticipated utilizing three twofold ML calculations are introduced in [4]. In this mistake investigation is finished with GLM, SVM and Bayes techniques. The precision of the basic ML techniques are analysed in [5] and variety in the exactness is given various sizes and information

divisions. The informational index of air quality comprises of poison information of CO, O3, NO2, SO2, PM10, and PM2.5. For the better air quality expectation, we should co-relate the toxin information with meteorological information [Temperature, Wind Speed, Humidity, Wind direction]. Neural organization strategy gives better exactness contrasted with others. The air contamination forecast is carried out in [6] based on various standard regularization and improvement calculations as an ML instruments.

For pollution estimation or prediction, linear regression algorithms are suitable and for forecasting the pollution levels neural network methods and SVM based methods are preferred [7].

The air quality file is anticipated by utilizing ML calculations for the recognition of PM2.5 level utilizing strategic relapse [8]. There are applications that show the steady PM2.5 levels, while some show the gauge of a particular day. This structure mishandles ML models to perceive and figure PM2.5 levels reliant upon an educational assortment comprising of meteorological conditions in a specific city. The informational index [9] utilized in discovery of PM2.5 level comprises of Temperature, Wind speed, Dew point, Pressure, PM2.5, Concentration (ug/m^3).

A. Gaps Identified in the Literature

In those papers, they only implemented the prediction of PM2.5. In this project they want to implement prediction of all the pollutants [CO, O3, NO2, SO2, PM2.5, PM10] with the help of meteorological data for better prediction.

III. EXPERIMENTAL ARRANGEMENT OF THE EXPECTED SYSTEM

A. Flow Chart The proposed technique is represented within the under-block diagram as proven in Fig. 1.

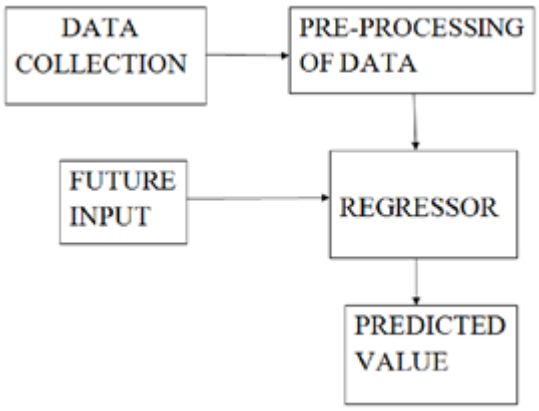

Fig. 1. Flow chart for the proposed approach

Central pollution control board built many pollution monitoring stations in heavily polluted areas, we collect the data from those monitoring stations.

B. Implementation of software

In Software specifications, used IDE is Anaconda python, Operating systems must be Windows 7/10 and we used the Coding language as Python.

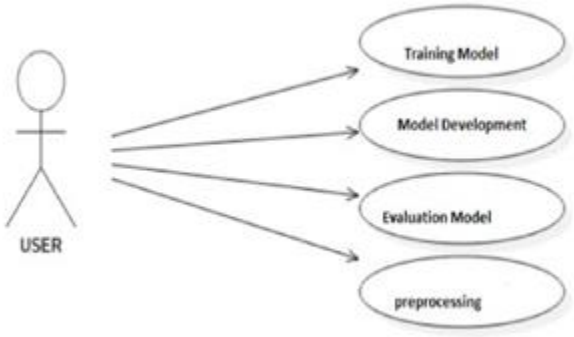

Fig. 2. Implementation flow of the model

IV. METHODOLOGY

A. Data Set

The Pollutant data:

Data information which is used to train the system to detect the air quality was obtained. The data set was to have attributes like CO, SO2, and O3.

Meteorological data:

The meteorological data information set parameters which is used to train the system are Temperature, Wind Speed, Humidity and Wind Direction.

B. USING MACHINE LEARNING MODELS

Linear regression:

Linear Regression [10] is nothing but an algorithm based on the machine learning are depends on supervised learning which performs a regression task. Depending on independent variables linear regression gives a target prediction value which is most likely used for finding the relationship among variables and forecasting. Depending on the connection among the established and the independent variables, different regression models differ, they are being considered and List of independent variables used.

$y = mx+c$

In the above expression y indicates labels to data and x indicates the input training data (input parameter).

Value of x is used to predict the value of y which gives best fit line for finding the best m and c values during training the model.

c = intercept m = slope of line

When we get the best m and c esteems, we get the best fit line. So when we are at long last utilizing our model for expectation, it will foresee the estimation of y for the information estimation of x.

Decision Tree:

The Regression on the Decision Tree [11] is both a non-linear and non-continuous construct. It represents a function that takes an attribute values vector as input, and returns a decision.

Decision tree falls within the Supervised Learning group. It can be used to solve regression as well as classification problems. By conducting a series of operations, a decision tree makes a decision.

Random Forest:

Natural forest is a method of bagging, and not of boosting. The trees are running in parallel in random woods. There is no contact among those trees while the trees are being installed.

It performs by constructing a multitude of decision trees during training time and outputting the class which is the particular trees ' class mode (classification) or average prediction(regression). A random forest [12] is a met estimator (i.e., it combines the outcome of many predictions) that aggregates many decision trees, with some useful improvements. The number of functions at each node that can be split on is limited to a certain percentage the total (known as the hyper parameter). Every tree takes a random sample from when, the original data set it generates its splits, adding another element of randomness which prevents over fitting.

Artificial Neural Networks:

The ANN system is a natural model, which is fascinating from information handling perspective since it computes and embraces choices and ends alike the human mind.

XGBoost regression:

Boost is a powerful approach for building supervised regression models. ... Ensemble learning involves training and combining individual models (known as base learners) to get a single prediction, and Boost is one of the ensembles learning methods

Lasso regression:

Lasso regression is a type of linear regression that uses shrinkage. Shrinkage is where data values are shrunk towards a central point, like the mean. The lasso procedure encourages simple, sparse models

V. RESULTS AND DISCUISION

We implemented the different machine learning algorithms in Python using Jupiter notebook. The following plot shows that all the features that are considered for the prediction are correlated and thus can be considered to train the model.

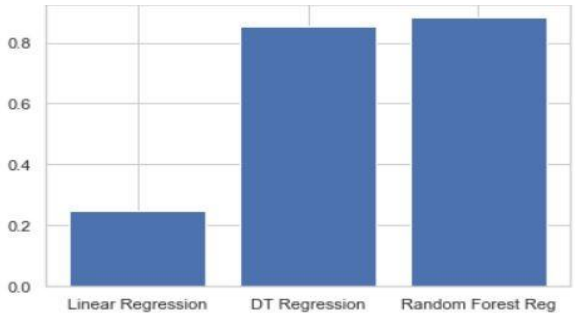

Fig. 3. SO2 prediction probability

For the SO2 prediction, we get prediction accuracy

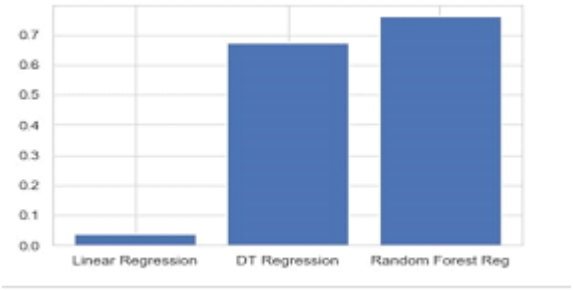

Fig. 4. CO prediction probability

For the CO prediction, we get prediction accuracy as follows

Type of Algorithm	Prediction Probability of CO
Linear Regression	0.02
Decision Tree	0.61
Random Forest regression	0.79

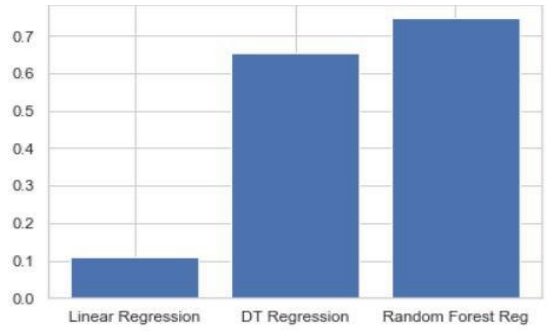

Fig. 5. O3 prediction probability

For the O3 prediction, we get prediction accuracy as follow

Type of Algorithm	Prediction Probability of O3
Linear Regression	0.09
Decision Tree	0.62
Random Forest regression	0.79

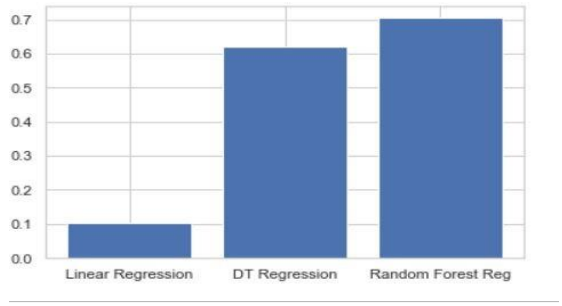

Fig. 6. NO2 prediction probability

For the NO2 prediction, we get prediction accuracy as follows

Type of Algorithm	Prediction Probability of NO2
Linear Regression	0.1
Decision Tree	0.64
Random Forest regression	0.701

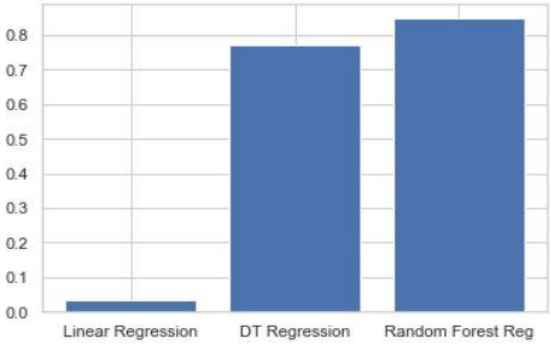

Fig. 7. PM2.5 prediction probability

For the PM2.5 prediction, we get mean square Coefficient as follows

Type of Algorithm	Prediction Probability of PM2.5
Linear Regression	0.02
Decision Tree	0.75
Random Forest regression	0.86
Type of Algorithm	**Prediction Probability of PM210**
Linear Regression	0.02
Decision Tree	0.61
Random Forest regression	0.79

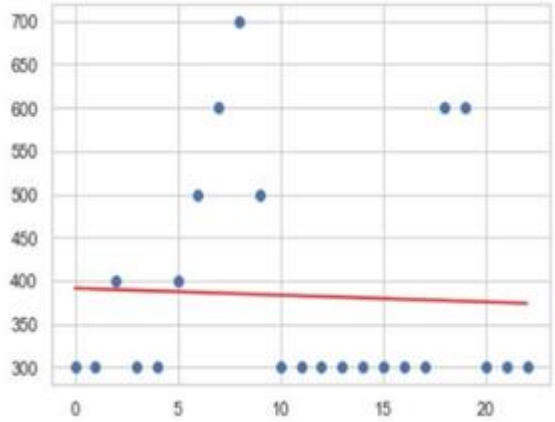

Fig. 8. Linear regression fitted curve for CO

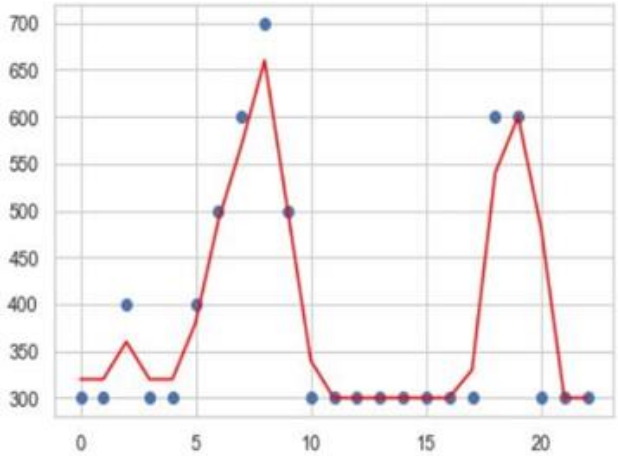

Fig. 9. Random forest fitted curve for CO

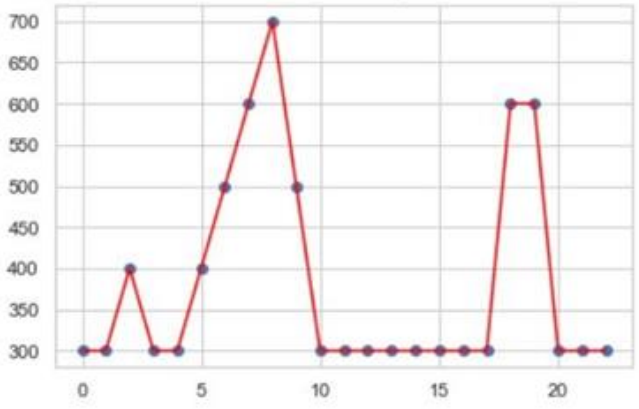

Fig. 10. Decision tree fitted curve for CO

VI. CONCLUSION

We anticipate the air quality list by utilizing distinctive calculations like direct relapse, Decision Tree and Random Forest. From the outcomes, we reasoned that the Random Forest calculation gives better expectation of air quality list.

VII. REFERENCES

[1]. https://en.wikipedia.org/wiki/Air_quality_index

[2]. Kennedy Okokpujie, Etinosa Noma-Osaghae, Odusami Modupe, Samuel John, and Oluga Oluwatosin, "A SMART AIR POLLUTION MONITORING SYSTEM," International Journal of Civil Engineering and Technology (IJCIET), vol. 9, no. 9, pp. 799–809, Sep. 2018.

[3]. Kostandina Veljanovska and Angel Dimoski, "Air Quality Index Prediction Using Simple Machine Learning Algorithms," International Journal of Emerging Trends & Technology in Computer Science, vol. 7, no. 1, 2018.

[4]. D. Zhu, C. Cai, T. Yang, and X. Zhou, "A Machine Learning Approach for Air Quality Prediction: Model Regularization and Optimization," Big Data and Cognitive Computing, vol. 2, no. 1, p. 5, Mar. 2018.

[5]. A. Masih, "Machine learning algorithms in air quality modeling," Global Journal of Environmental Science and Management, vol. 5, no. 4, pp. 515–534, 2019.

[6]. https://archive.ics.uci.edu/ml/datasets/Air+quality

[7]. Rokach, Lior; Maimon, O. (2008). Data mining with decision trees : theory and applications. World Scientific Pub Co Inc. ISBN 978-9812771711.

[8]. BreimanL (2001)."RandomForests". MachineLearning. 45 (1):32. doi:10.1023/A:1010933404324

Cite this article as :

Ritik Sharma, Gaurav Shilimkar, Shivam Pisal, " Air Quality Prediction by Machine Learning", International Journal of Scientific Research in Science and Technology(IJSRST), Print ISSN : 2395-6011, Online ISSN : 2395-602X,Volume 8, Issue 3, pp.486-492, May-June-2021. Available at doi : https://doi.org/10.32628/IJSRST218396
Journal URL : https://ijsrst.com/IJSRST218396