IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

Impact Factor 8.021 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 12, Issue 4, April 2024

DOI: 10.17148/IJIREEICE.2024.12405

Machine Learning Based Air pollution Predictor

Sridevi N[1], PrasannaKumar M[2]

Department of Electronics and Instrumentation Engineering, Dr. Ambedkar Institute of Tecnology, Bangalore, Karantaka, India. [1,2]

Abstract: Over last few decades' pollution in the air have been increasing promptly over the globe, due to growing urbanization and industrialization. The degrading air quality effects the climate condition and hence severe health issues related to lung and respiratory system. Hence, monitoring the quality of air becomes crucial in keeping the atmosphere and global ecosystem clean. However, the existing methods such as Probability and Statistics are more complex method to detect the air quality. In order to overcome from the above issues this paper proposes a system to monitor the air pollution using Machine Learning. A new Dataset is created to by measuring different pollutants in air arund Nagarbhavi area in Bengaluru, Karnataka, India using sensors. To validate the method the actual value is compared with predicted value.

Keywords: Air Quality, Prediction, Machine Learning, classification, dataset

I. INTRODUCTION

Today, one amoung the utmost significant environmental health hazard faced globally is air pollution. Which is caused by both human and natural sources [1]. Natural sources can also pollute the air like methane produced by decomposing organic material in soil. Gases and ash from volcanic outbreaks etc. Moreover, common pollutants are created by human like emissions produced by factories and vehicles, smoke from cigarette and by burning matters etc. Air pollution has become a significant environmental concern, affecting the health and welfare of individuals and the entire feature of life in many regions. Hence, it is crucial to monitor air pollution levels in real-time and predict future pollution levels to mitigate its adverse effects and implement appropriate measures for pollution control. Air Quality Index (AQI), measures the quality of atmospheric air, speed of the wind, direction of the wind, relative humidity and temperature. Linear Regression (LR), SVM, Decision Tree (DT), Random Forest (RF), KNN technique was applied in [2][3] to detect the quality of air.

Experiment was conducted in [4] on three different regions to validate the algorithm such as The Support Vector Machine (SVM), Artificial neural network (ANN), AdaBoost, Random Forest (RF) and Stacking Ensemble. Various statistical methods are used to estimate the AQI. The absorption of SO2 in air was analysed in [5], here, in this paper Auto Regressive (AR) and ARIMA technique was adopted to detect the degradation of environment due to SO2. Reference [6] explains the different ML techniques like LR, RF, KNN and SVM for predicting the air quality. Prediction of pollutant in air was discussed in [7] using neural network and SVM in Delhi, India using CPCB dataset. Authors in [8] reviewed the different pollutants and methods used for predicting the AQI. $PM_{2.5}$ has highest influence in degrading air the quality. Here, neural network, regression models and classification techniques are discussed. In the latest year, air pollution pays to 11.65% of death across the world. Prolong explosion to air pollution creates severe health related issues like COPD, lung cancer, lung infection, etc

Organization: The organization of the remaining topics of this article is: Section II explains the materials and different methods adopted in this work, followed by performance analysis in section III and finally conclusions are drawn in section IV.

II. SYSTEM DESIGN

The entire work has been divided into two parts consisting of predicting AQI from existing dataset and creating dataset using data collected from sensor from air. To be able to gather data for the construction and training of the model to be used in the machine learning algorithm, a prototype is build as a air quality monitoring system, which consisting of an integrated array of sensors, an Arduino microcontroller powered by SMPS. The sensors used in this work are MQ2, MQ3, MQ9, and MQ135 gas sensors. The air pollutants which considered in this research are alcohol, smoke, carbon monoxide and overall air quality. Overall air quality is affected by various gases like, air pollution gases (e.g., CO, NO, SO_2 etc.) and greenhouse gases (e.g., CO_2, CH_4, N_xO). Alcohol indirectly is a main contributor for air pollution, i.e., which helps in formation of smog.

IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

Impact Factor 8.021 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 12, Issue 4, April 2024

DOI: 10.17148/IJIREEICE.2024.12405

Smoke primarily consists of particles and can include other gaseous air pollutants, like nitrogen oxides, carbon monoxide, and hydrocarbons. Table below summarizes the sensors used in constructing the hardware. Table 1 illustrates the senosrs and pullutants details.

Table 1: The details of Gas sensors and Pollutants detected

Gas Sensor	Pollutant Detected	Gas Sensor	Pollutant Detected
MQ-2	Smoke	MQ-9	Carbon Monoxide
MQ-3	Alcohol	MQ-135	General Air Quality

Further in the prediction part, ML technique helps to predict AQI. There are various ML algorithm amongst K-Nearest Neighbor. KNN is sultable for regression applications, this algorithm predicts the future values of pollutants. Initially, the data is gathered from CPCB website which helped to give some insight about the partical matters in air.

The steps applied for prediction is shown in figure 2.

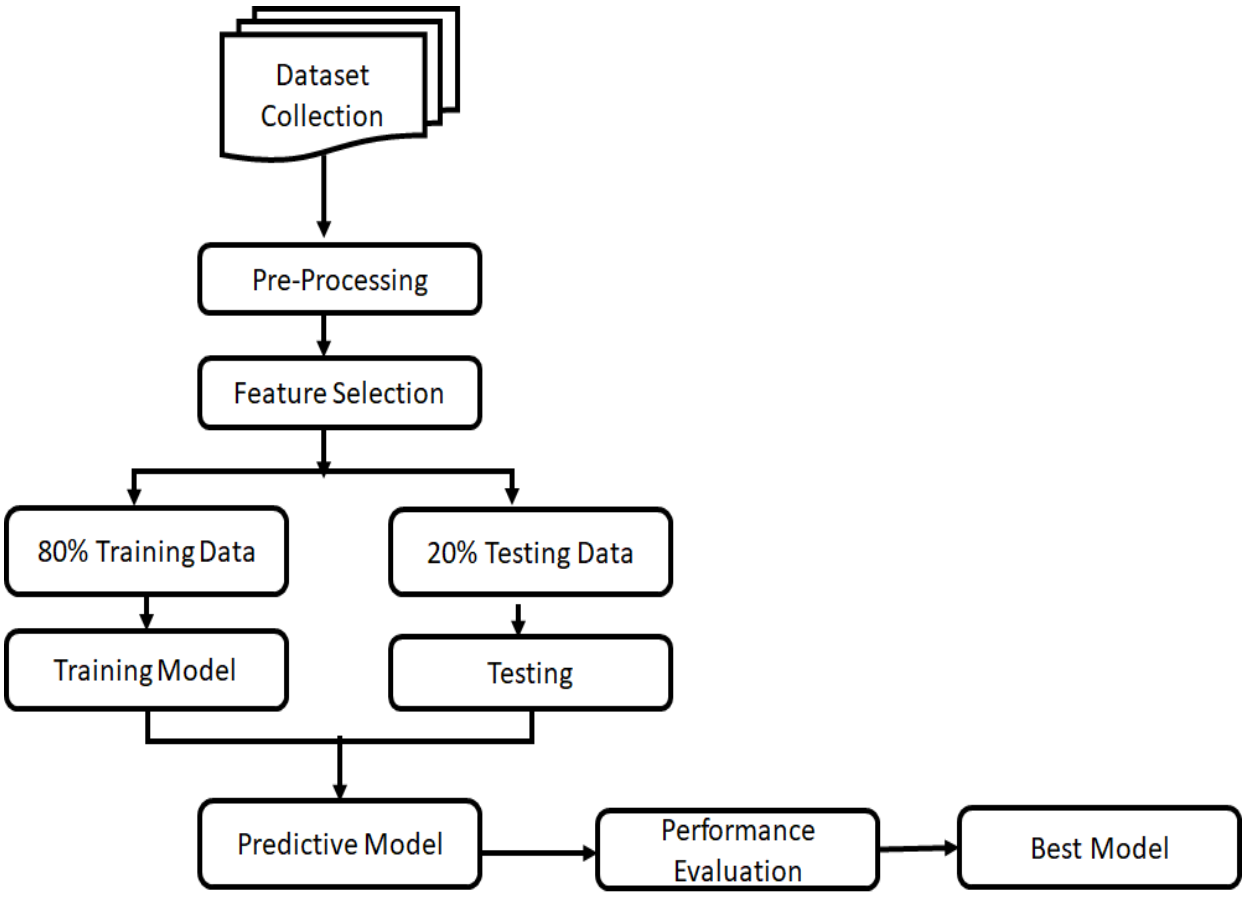

Fig. 2. Process flow of AQI prediction

A. DATA COLLECTION:

The dataset from central pollution and control board government of India is chosen for study purpose. The dataset used in this work contains 12 parameters such as SO2, NO2, SPM and RSPM of various cities. Further, from the data set the observations of Karnataka state is chosen for analysis. There is a prescribed index values for Air Quality Index (AQI) used by Central Pollution Control Board (CPCB) which are given below in table 1.

IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

Impact Factor 8.021 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 12, Issue 4, April 2024

DOI: 10.17148/IJIREEICE.2024.12405

Table 2: The ranges of air quality index and categories

AIR QUALITY INDEX(AQI)	CATEGORY
0-50	Good
51-100	Satisfactory
101-200	Moderate
201-300	Poor
301-400	Very Poor
401-500	Severe

B. Data Pre-Processing:

ML model requires the data to be inputed in a specified format to accomplish better results. In this step, the raw data extracted from dataset is prepared and make it to suitable for ML model.

The collected dataset should be of CSV (Comma Separated Values) format, with comma delimiter. AQI data of Bangalore city from Karnataka is extracted and the NaN values of the data are filled using mean value. The sample dataset is tabulated in table 2 for the year 2015. The dataset includes parameters like so2, no2, rspm, spm.

There are few data pre-processing steps. The dataset should be of CSV (Comma Separated Values) format, with comma delimiter. It includes to get basic summary about the dataset being used, i.e., information signifying the number of rows or columns exist in the dataset, checking for 'NaN' values or any other factors. It uses methods like shape and info for it. Using drop method, it can remove the specific rows or columns. The dataset is then fitted to predict the values.

C. Air Quality Index (AQI) Calculation:

The prediction part is done using supervised ML algorithm. The algorithm used is K-Nearest Neighbor (KNN). Choosing the 'k' parameter is very important. It is done based on the underlying structure and features present in the dataset. A small value of 'k' indicates higher influence of noise on the outcome and a greater value of 'k' can make the computation expensive.

The chosen 'k' also needs to produce minimum error rate. The 'k' value defines how many neighbors will be checked to determine the classification of a specific query point. For this a distance metric is used. The most common distance metric used is Euclidean Distance. Then, it assigns the point to the class among its k nearest neighbor. Below is the depiction for Euclidean Distance.

The flowchart in figure 3 shows the working mechanism of the circuit. The power supply needs to be switched on for the circuit to start up. Then the OLED displays the initialized parameters. Then all the sensors start detecting their respective parameters, the raw values are sent to the Arduino Microcontroller.

This microcontroller converts analog to digital values using ADC. Then the value is checked whether it is within the set threshold value (set in the code written in Arduino IDE) or not. If it is not within the threshold value, then, a warning message displayed along a buzzer to notify the detected pollutants on the OLED.

IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

Impact Factor 8.021 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 12, Issue 4, April 2024

DOI: 10.17148/IJIREEICE.2024.12405

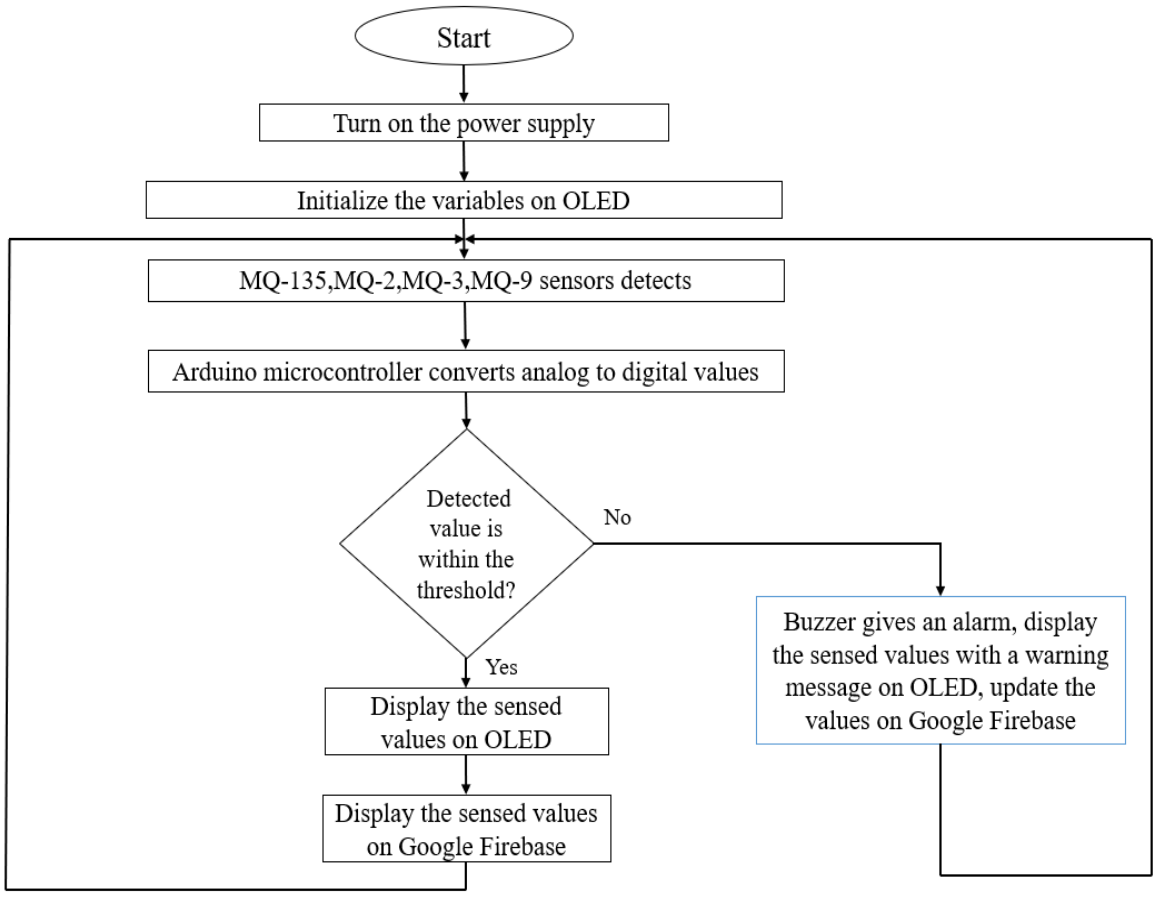

Fig.3. Working mechanism

III. EXPERIMENTAL ANALYSIS

The designed method is validated by conducting measuremnets using various evaluation metrics like accuracy score, precision score, recall, score and F1 score. The base for these metrics is a Confusion Matrix. A confusion matrix is a matrix that is used to evaluate the performance of a machine learning model. It is used to show the number of true positives, false positives, true negatives and false negatives that the model has produced for a given dataset. True Positives (TP), False Positives (FP), True Negatives (TN), and False Negatives (FN) are the four possible outcomes of a binary classification problem. As there are multiple classes in our dataset, the confusion matrix will produce more set of values. The TP, TN, FP, FN values are needed to be computed for each class. Then the average value of each metric is taken to be the final value.

The several estimation metrics used are described below

1. Accuracy Score: It measures the quantity of correct predictions made by the model. Hence, the calculation done as the ratio of the number of correct predictions to the total number of predictions. The formula for accuracy score is

$$Accuracy = \frac{true\ positives + true\ negatives}{true\ positives + true\ negatives + false\ positives + false\ negatives}$$

2. Precision: It measures the quantity of true positive predictions made by the model. It is calculated as the ratio of the number of true positives to the sum of true positives and false positives. The formula for precision score is

$$precision = \frac{true\ positives}{true\ positives + false\ positives}$$

IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

Impact Factor 8.021 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 12, Issue 4, April 2024

DOI: 10.17148/IJIREEICE.2024.12405

3. Recall: It measures the quantity of true positive predictions made by the model out of all the actual positive samples. It is calculated as the ratio of the number of true positives to the sum of true positives and false negatives. The formula is given below

$$recall = \frac{true\ \ positives}{true\ positives + false\ negatives}$$

4. F1 Score: A weighted average of precision and recall. It is calculated as shown below

$$F1\ Score\ =\ 2\ *\ ((Precision\ *\ Recall)\ /\ (Precision\ +\ Recall))$$

The different parmeters calculated to validate the method is listed in table 3

Table 3: Performance Metrics

Accuracy Score	Precision	Recall	F1 Score
0.8297	0.6523	0.5946	0.6219

Fig 4 shows the scatter plot to know how much one variable is affected by another or the association between them in two dimensions. The measured and predicted values are listed in table 4. Similarly the comparison between real values and predicted values are illustrated in figure 5.

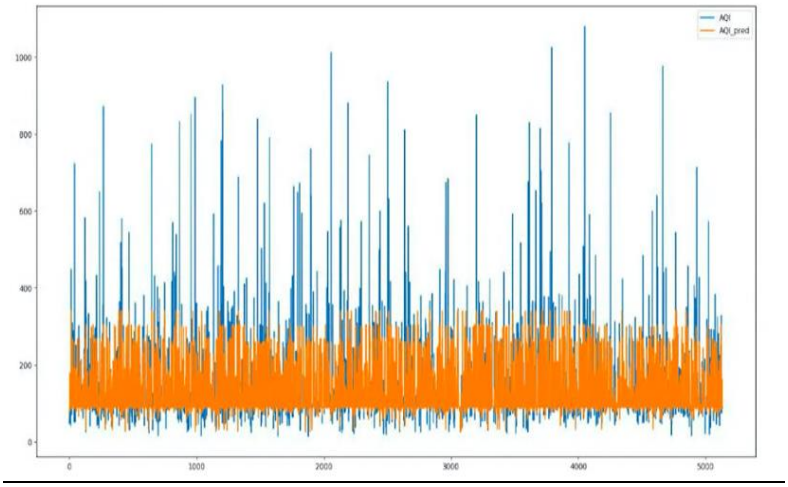

Fig 4. Graphical representation for the values plotted between AQI and AQI_pred

Table 4: Actual and predicted values

	A	B	C	D	E	F	G
1	Location		1	2	3	4	5
2	Actual	MQ-2	0.26	0.18	0.12	0.14	0.14
3		MQ-3	1.43	1.13	0.7	0.58	0.63
4		MQ-9	1.6	1	0.95	0.72	0.74
5		MQ-135	1.1	0.84	0.48	0.46	0.7
6							
7	Predicted	MQ-2	0.27	0.25	0.11	0.08	0.27
8		MQ-3	1.65	1.43	0.7	0.13	0.52
9		MQ-9	2.54	1.6	0.65	0.16	0.62
10		MQ-135	1.45	1.15	0.43	0.58	0.85
11							

IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

Impact Factor 8.021 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 12, Issue 4, April 2024

DOI: 10.17148/IJIREEICE.2024.12405

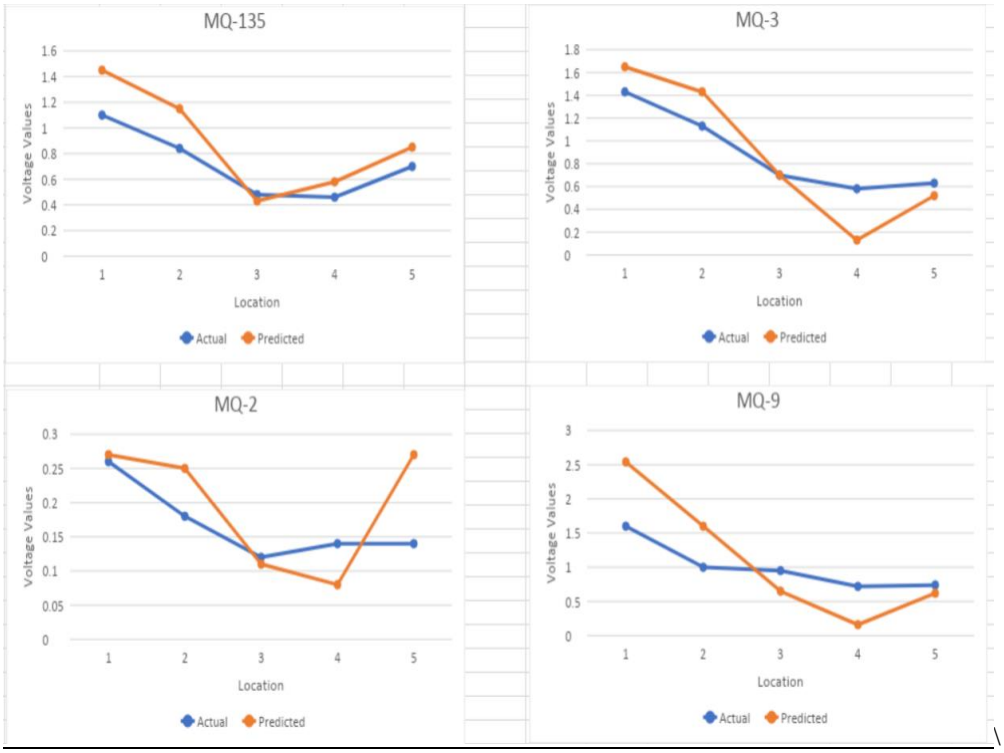

Fig 5. The comparison of actual and predicted values

IV. CONCLUSION

In this work, the concentration of air pollutants such as Carbon Monoxide, Smoke, and Alcohol in ambient air is administered and the levels of each gas is recorded. ML approach KNN is used to predict AQI. Also daset is created by measuring the pollutants using different sensors. To validate the approach the actual value is compared with predicted value. Further, different metrics are calculated to demonstrate the algorithms performance.

REFERENCES

[1]. Zhang, C.; Yan, J.; Li, Y.; Sun, F.; Yan, J.; Zhang, D.; Rui, X.; Bie, R. Early air pollution forecasting as a service: An ensemble learning approach. In Proceedings of the 2017 IEEE International Conference on Web Services (ICWS), Honolulu, HI, USA, 25–30 June 2017; pp. 636–643.

[2]. T. Madan, S. Sagar and D. Virmani, "Air Quality Prediction using Machine Learning Algorithms –A Review," *2020 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN)*, Greater Noida, India, 2020, pp. 140-145, doi: 10.1109/ICACCCN51052.2020.9362912.

[3] Gaganjot Kaur, Jerry Gao, Sen Chiao, Shengqiang Lu, Gang Xie, Air Quality Prediction: Big Data and Machine Learning Approaches, International Journal of Environmental Science and Development 9(1):8-16, DOI: 10.18178/ijesd.2018.9.1.1066

[4]. Yun-Chia Liang ,Yona Maimury ,Angela Hsiang-Ling Chen andJosue Rodolfo Cuevas Juarez, Machine Learning-Based Prediction of Air Quality, *MDPI, Appl. Sci.* 2020, *10*(24), 9151, https://doi.org/10.3390/app10249151.

[5]. Pooja Bhalgat, Sejal Pitale, Sachin Bhoite, Air Quality Prediction using Machine Learning Algorithm, International Journal of Computer Applications Technology and Research Volume 8–Issue 09, 367-370, 2019, ISSN:-2319–865

[6]. K. Mahesh Babu, J. Rene Beulah, Air Quality Prediction based on Supervised Machine Learning Methods, International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-8, Issue-9S4, July 2019

[7]. U. Mahalingam, K. Elangovan, H. Dobhal, C. Valliappa, S. Shrestha and G. Kedam, "A Machine Learning Model for Air Quality Prediction for Smart Cities," *2019 International Conference on Wireless Communications Signal Processing and Networking (WiSPNET)*, Chennai, India, 2019, pp. 452-457, doi: 10.1109/WiSPNET45539.2019.9032734.

[8]. Iskandaryan, D.; Ramos, F.; Trilles, S. Air Quality Prediction in Smart Cities Using Machine Learning Technologies Based on Sensor Data: A Review. *Appl. Sci.* 2020, *10*, 2401. https://doi.org/10.3390/app10072401