JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

PREDICTION AND ANALYZING AIR QUALITY OF SHIROLI MIDC USING MACHINE LEARNING

[1]ANAGHA G. PATIL, [2]SHRIKANT M.BHOSALE, [3]RASHMI J. DESHMUKH,[4] SHEETAL GAIKWAD

[1,2]Enviromental science and technology Engineering Department, [3,4]Computer science Department.

Abstract : As cities grow both financially and technologically, pollution problems such as noise,water and air pollution become more prevalent. Particularly poor air quality has an effect on people's health. Fossil energy usage, rapidly expanding industrialisation, automobile exhaust, etc. are the main sources of air pollution. Waste generation represents one of the most significant contributors to air pollution. Carbon Monoxide, nitrogen Dioxide (No2) and Sulfur Dioxide (So2) are the major components of pollution of air. Various air quality surveillance devices keep an eye on various pollutants. Because the amount of contaminants in the air changes throughout the day, forecasting the air quality is challenging. Various Machine learning algorithms can be used to solve this issue. With the help of machine learning (ML) techniques and IoT sensors are installed at different stations of Shiroli located in Maharashtra Industrial Development Corporation District Kolhapur, Maharashtra. Data is collected hourly and preprocessed. The data obtained by industrial area are preprocessed and relevant features are extracted using different ML algorithm. Results of the algorithms are compared and performance is evaluated with different measures like R, R2 MAE, RMSE, etc.

Keywords: Air pollution, IoT sensors, Machine learning, AQI etc.

1. INTRODUCTION

Environmental pollution such as noise, water and air pollution are increasing as development of cities rapidly increasing economically and technologically. Air pollution, In particular, Air pollution has a severe impact on human health. Pollution and particle matter which has piqued people's interest. Among the scientific community, there is a growing interest in air pollution and its consequences. The main causes of air pollution are fossil fuel combustion, agriculture, and emissions from vehicles. Manufacturing and businesses, domestic heating, and natural gas are only a few examples.

The environment is defined as everything that surrounds us. The environment is being polluted as a result of human activities and natural disasters, the most serious of which is air pollution. When the humidity is high, we feel significantly hotter because moisture does not escape into the air. As a result of expanded transportation infrastructure, industrialization is one of the major contributors to air pollution. Another significant contributor to air pollution is industrialization. Particulate matter (PM), SO2, Nitrogen Oxide, Carbon Monoxide and other significant pollutants insufficient oxidation of gas and oil results in the production of carbon monoxide.

When thermal fuel is ignited, oxygen is produced. Headaches and vomiting are caused by carbon monoxide. Because benzene is produced as a result of smoking, it causes respiratory problems. Dizziness and nausea are caused by nitrogen oxides. Particulate matter with a diameter of 2.5 micrometers or less has a greater impact on health of people. There must be actions taken to lessen environmental air pollution. To evaluate the quality of the air, one uses the Air Quality Index (AQI). Prior to now, the capacity to estimate air quality relied on traditional approaches like probability and statistics, but these approaches are exceedingly difficult to utilize. Because of technological advancements, it is now very simple to obtain data about air pollutants using sensors.

The air quality forecasting models

Forecasting the air quality can help society and the government makes informed decisions about the environment and can show pollution patterns in advance. The time-scale categorization of forecasting models, forecasting methodology, and input type are some fundamental indices for the approximate classification of air quality forecasting models. Air quality prediction contains categories for the very short-term, short-term, medium-term, and long-term based on the data resolution of time. The studies under consideration primarily use hourly and daily resolution for their predicting data. The daily readings often show long-term patterns in the development of contamination and are derived from data collected over years. In general, a shorter forecasting temporal extent can produce results that are more precise and in-depth, while a larger forecasting temporal extent can supply research with long-term data. Hourly prediction could provide more accurate short-term air quality monitoring and management, while this type of application of prediction research aids in the realization of long-term pollution control operations. The process of setting up air quality modeling is a difficult part of environmental science research and a complicated part of system engineering. Through detailed investigation, it aids in the research of the significance of the causes and consequences of contaminants and contributes to potential future mitigation measures. Air quality forecasting models have undergone continuous development over the past few years so that a variety of approaches and methods for time series data can be used to target pollution control and avoid serious pollution incidents.

Recent studies focus on the use of sophisticated statistical learning techniques for the assessment of air quality and the prediction of air pollution. Now days for accurate prediction of AQI model are done using Machine Learning (ML).

2. LITERATURE REVIEW

Mauro Castelli *et.al* (2020) studied A Machine Learning Approach to Predict Air Quality in California. Support vector regression was used to forecast pollution and particle levels and properly identify the AQI in this investigation. They created a model of hourly atmospheric pollution that enabled them to simulate pollutant concentrations including O3, CO, and SO2 along with the hourly AQI with a high degree of accuracy.

Madhuri VM and Samyama Gunjal GH (2020) studied Air Pollution Prediction Using Machine Learning Supervised Learning Approach. The amount of air pollutants in ambient air depends on meteorological factors such atmospheric wind speed, wind direction, relative humidity, and temperature. A tool for evaluating air quality is the Air Quality Index (AQI). The proposed study uses the LR, SVM, DT, and RF algorithms as part of a supervised learning method.

Dixian Zhu *et.al* (2018) studied A Machine Learning Approach for Air Quality Prediction: Regularization and Optimization of Models. In this study, they developed effective machine learning methods for predicting air pollution. They solved many formulations of the issue using cutting-edge optimization techniques, which they then formalized as regularized MTL. They focused on improving performance by using a structured regularized while reducing model complexity by reducing the number of model parameters. According to their findings, The suggested light formulation performs better than conventional model formulations, and regularization, which entails requiring prediction models to remain near over a period of two hours, can also improve forecast precision.

Mrs. A. Gnana Soundari *et.al* (2019) studied an Indian Air Quality Prediction and Analysis using Machine Learning. In order to forecast India's air quality, they employed machine learning to calculate the area's air quality index. They developed a model that can correctly estimate the future air quality index of any data inside a certain region with 95% accuracy and can anticipate current data. They used a model to predict the AQI.

Yun-Chia Liang *et.al* (2020) studied Machine Learning-Based Prediction of Air Quality. For AQI forecasting, using artificial intelligence approaches yields promising results. The Taiwanese EPA and CWB gathered data for this study during an 11-year period. This study also demonstrates that different Taiwanese regions have different prediction performance. When results from datasets collected from three different regions are compared, the results for Fengshan AQI prediction are the most accurate (Southern Taiwan). For forecasts of one hour, eight hours, and twenty four hours, respectively, 95 percent confidence intervals are produced. A decision-maker may find the 95 percent C.I. more useful as a guide than a single value estimate.

Avijoy chakma *et.al* (2020) studied Image-based air quality analysis using deep convolutional neural network. Existing image-based haze level analysis systems are mostly inspired by the dazing techniques. They present the first study in the literature that makes use of a CNN-based method to calculate PM2.5 concentration for images taken naturally. Using two transfer learning techniques—CNN fine tuning and CNN feature-based Random Forest—the images are divided into three groups based on their PM2.5 concentrations.

3. METHODOLOGY

Selection of study area.

The selected area for Air Quality analysis is Shiroli (MIDC). This industrial area is located near Kolhapur city in the state of Maharashtra, India. The total area of the industry is 254.19 Hectare.

The main industries present in the Shiroli (MIDC) area are

- Agricultural products
- Mechanical items
- Chemical items
- Mechanical items
- Casting
- Packaging
- Food items
- Glass and Ceramic industry

Air quality analysis using IoT/HVS:

The various IoT sensors and different Laboratory Instruments will be used for the Air quality analysis of selected area.

Different sensors used for calculating CO, NO_2 and So_2. The arduino based sensors are used for analyzing air quality.

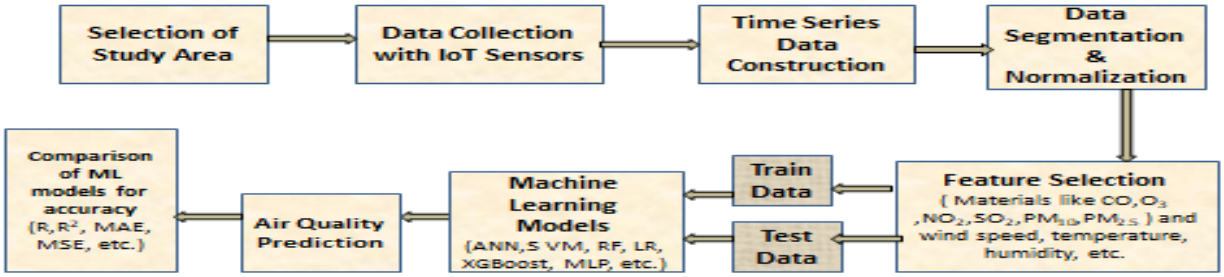

Fig 1.flow chart of methodology sampling sites and sampling frequency

Sensors or the lab Instrument will located at 2 or more stations in the area

1) TRIO INDUSTRY
2) QUALITY FABRICATIONS

Recording the data from sensors using IOT and other air quality measuring instruments.

The frequency of data collection from station will be daily basis .The readings are taken at hourly basis.

Detection of PM, CO_2, SO_2, NO_2 etc. by using IoT sensors.

Data collection and Data processing with the help of machine learning tools.

Data is collected in hourly basis and then this data is process with the help of various machine learning tools and algorithms. For prediction of AQI model is done using Machine Learning (ML).The study of machine learning algorithms for index calculation. i.e.M5P, random tree, random forest, support vector, Gaussian, artificial neural network.

Prediction and analyzing the collected data using machine learning.

4. DETAILS EXPERIMENTAL

The selected area for Air Quality analysis is Shiroli (MIDC). This industrial area is located near Kolhapur city in the state of Maharashtra, India. The total area of the industry is 254.19 Hectare. The main industries present in the Shiroli (MIDC) area are - Agricultural products, Mechanical items, Chemical items, Mechanical items, Casting, Packaging , Food items, Glass and Ceramic industry etc.

The various IoT sensors and different Laboratory Instruments will be used for the Air quality analysis of selected area.

Different sensors used for calculating CO, NO_2 and So_2. The arduino based sensors are used for analyzing air quality.

Sr. No	Pollutant	Sensor
1	Carbon Monoxide (CO)	MQ7-BB
2	Nitrogen Dioxide (NO_2)	SPEC Sensor LLC 110-507
3	Sulfur Dioxide (SO_2)	SPEC Sensor LLC 110-602
4	Particular Matter 2.5 (PM2.5)	DERobot Gravity

Table 1 Sensor list

Data Preprocessing in Machine learning

Data preparation is the process of transforming raw data into something a machine learning model can use. It is the initial and most crucial step in the process of creating a model for machine learning

AQI calculation and classification.

$$IP = \frac{IHI - ILO}{BPHI - BPLO}(CP - BPLO) + ILO$$

I_P= the index for pollutant p
BP_{HI}=concentration breakpoint that is higher than C_p
BP_{LO}=concentration breakpoint that is lesser thanC_p
I_{HI} =AQI value referring to BP_{HI}
I_{LO} =AQI value referring to BP_{LO}
C_P= Input concentration of given pollutant

Classification of air quality using AQI.

Air Quality Index Value	Air Quality	Numbering with respect to Air Quality
0 to 50	Good	0
51 to 100	Satisfactory	1
101 to 150	Moderate	2
151 to 200	Poor	3
201 to 300	Very Poor	4
301 to 500	Severe	5

Table .2 AQI classification

5. RESULT

Result obtained by WEKA software.
Total number of instances consider for calculation are 372

Algorithm used	Correlation coefficient	MAE	RMSE
Decision Table	1	0	0
Gaussian	0.9209	0.1538	0.19
M5P	0.9993	0.016	0.0207
Simple LR	0.9148	0.1533	0.1968
Linear Regression	0.9329	0.1397	0.1755
Random Forest	0.9997	0.0014	0.0112
Random tree	0.9943	0.0027	0.0518

Table 3 Algorithm results

Mean Absolute Error (MAE)

The MAE in Decision Tree algorithm is nearly zero and the error value increases as we go to Random Forest, Random Tree, M5P, Linear Regression, Simple LR and Gaussian respectively.

Root mean squared error

The RMSE in Decision Tree algorithm is nearly zero and the error value increases as we go to Random Forest, M5P, Random Tree, Linear Regression, Gaussian and Simple LR respectively.

Result obtained by Random Tree algorithm.

The value on the line joining parent node is representing the splitting criteria based on the AQI value based on CO concentration in air. The first value on this leaf represent the AQI value based on Co concentration is less than 303.5 is correctly classified in 144 instances out of 372.The second value on the node represent AQI value based on Co concentration greater than 303.5 incorrectly classified in 228 instances out of 372.

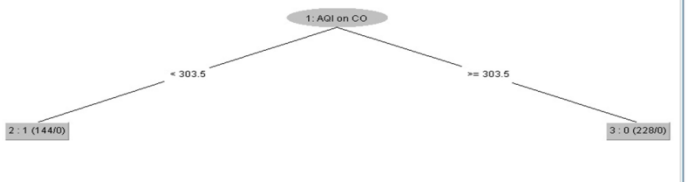

Fig. 2 Results from RT algorithm

Result obtained by M5P algorithm.

The value on the line joining parent node is representing the splitting criteria based on the CO concentration in air. The first value on this leaf represent the concentration of Co is less than 17.5 mg/m^3 is correctly classified in 144 instances out of 372.The second value on the node represent Co concentration greater than 17.5 mg/m3 incorrectly classified in 228 instances out of 372.

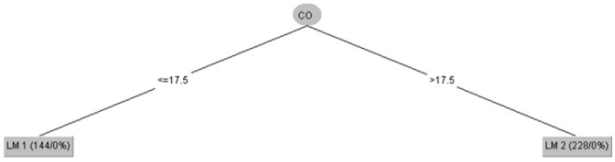

Fig.3 Results from M5P algorithm

Result obtained by Jupyter Software

Correlation Matrix

A table displaying correlation coefficients between variables is called a correlation matrix. The correlation between two variables is displayed in each cell of the table. Data are summarized, input into more complex studies, and diagnostics for complex analyses.

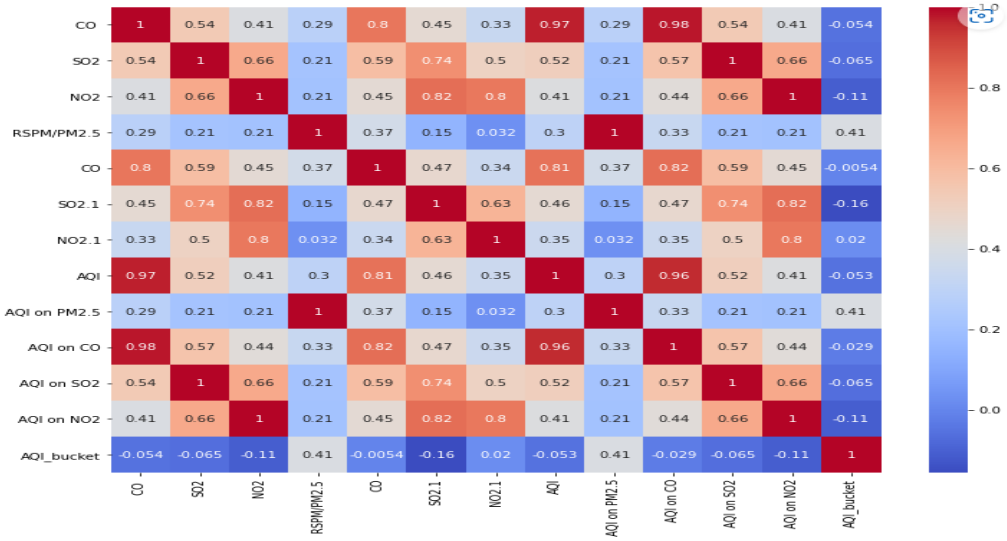

Fig.4 correlation matrix

The result obtained after data processing with Machine Learning (ML)

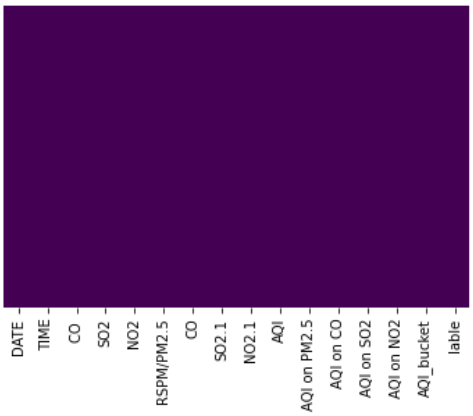

Fig.5ML graph

The result obtained after data processing with Machine Learning (ML)

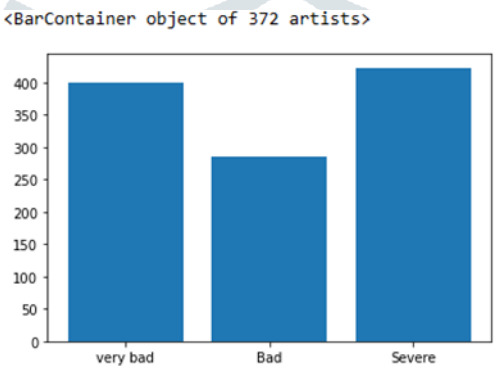

Fig.6 Histogram

6. CONCLUSION

Any substance that disturbs the natural or original properties of the air whether it be chemically, physically or biologically is considered an air pollutant .Industrial pollution is one of the main sources of pollution in the globe is this type of pollution. AQI is the index used to calculate daily air quality. AQI gives information about the air quality and also gives the information about potential negative health impacts. IoT sensors are used to calculate pollutant concentration in the air. IoT sensors increases accuracy minimizes the cost and reduces effort.

WEKA tool plays a vital role in predicting AQI with classification algorithm. We have experimented monthly data of Industrial area of Shiroli (MIDC) located near Kolhapur city for experimentation. For analysis various class algorithms are used including LR, SVM, DT, RT, M5P, Simple LR, Gaussian and RF. The different algorithm compare with different air quality parameter like Correlation coefficient, Mean absolute error, Relative absolute error, Root relative squared error and Total no of Instances.

Decision Tree is most suitable algorithm compare to other DT gives least MAE and RMSE. Total number of instances taken for experimentation is 372.The Mean Absolute Error and Root Mean Square Error for Decision Tree algorithm is nearly 0. With the help of this algorithm, we are able to create an accurate model of the industrial air pollution and achieve usually high accuracy. Different algorithms like Random Forest, M5P and Random Tree show the promising results and error values lies between 0 to 0.2.

REFERENCES

1. Sujuan Liu, Chuyu Xia, Zhenzhen Zhao (2016) “A low power real time air quality monitoring system using LPWAN” based on LoRa. 13thIEEE(ICSICT)379-381
2. R. du Plessis, A. Kumar, G. Hancke and B. Silva(2016)"A wireless system for indoor air quality monitoring," IECON 2016 - 42nd Annual Conference of the IEEE Industrial Electronics Society, 2016, pp. 5409-5414
3. Mauro Castelli , Fabiana Martins Clemente,Aleš Popovič,1, Sara Silva,3 and Leonardo Vanneschi(2020) “A Machine Learning Approach to Predict Air Quality in California” Hindawi Complexit Volume 2020, Article ID 8049504, 23 pages
4. Madhuri VM, Samyama Gunjal GH, Savitha Kamalapurkar(2020) “Air Pollution Prediction Using Machine Learning Supervised Learning Approach” International Journal of Scientific & Technology Research volume 9
5. Dixian Zhu1, Changjie Cai2 , Tianbao Yang and Xun Zhou(2018) “A Machine Learning Approach for Air Quality Prediction: Model Regularization and Optimization” Big Data Cogn. Comput.
6. Mrs. A. Gnana Soundari, Mrs. J. Gnana Jeslin M.E, Akshaya A.C(2019) “Indian Air Quality Prediction And Analysis Using Machine Learning” International Journal of Applied Engineering Research ISSN 0973-4562 Volume 14,

7. Yun-Chia Liang1, Yona Maimury , Angela Hsiang-Ling Chen2, and Josue Rodolfo Cuevas Juarez(2020) "Machine Learning-Based Prediction of Air Quality" Appl. Sci. 2020, 10, 9151
8. Lindgren and Saravanakumar, R.(2009). "Air pollution tolerance index of selected plants" Asian Journal of Environment Science. 6:161
9. Mage, D., Ozolins G., Peterson, P., Webster A,Orthofer,R.Wandeweerd V., and Gwynne, M. (1996). "Urban air pollution in megacities of the world" Atmospheric Environ., 30: 681.
10. Gupta, M.C. and Ghouse, A.K.M. (1987)."The effect of coal smoke pollutanatnts on the growth ,yield and leaf epidermal features of Abelmoschus esculentus moench" Environ. Pollut., 47: 221.
11. Indian Standard. (1975) "Method for measurement of air pollution - nitrogen dioxide". IS 5182 (Part 6)
12. Indian Standard. (2001) "Method for measurement of air pollution - sulphur dioxide". IS 5182 (Part 2)
13. Yu, L.; Wang, S.; Lai, K.K.(2007) "Basic Learning Principles of Artificial Neural Networks In Foreign-Exchange-Rate Forecasting With Artificial Neural Networks" Springer: Boston, MA, USA, 2007
14. Verma, Ishan, Rahul Ahuja, HardikMeisheri, and LipikaDey. "Air pollutant severity rediction using Bi-directional LSTM Network" In 2018 IEEE/WIC/ACM International
15. Zhang, Chao, Baoxian Liu, Junchi Yan,Jinghai Yan, Lingjun Li, Dawei Zhang, XiaoguangRui and RongfangBie. "Hybrid Measurement of Air Quality as a 5 Fig. 8. RH w.r.t tin oxide Fig. 9. RH w.r.t C6H6 Mobile Service: An Image Based Approach." In 2017 IEEE International Conference on Web Services (ICWS), pp. 853- 856. IEEE,2017.
16. Y. Zhou, X. Zhao, K.-P. Lin, C.-H. Wang, L. Li, A Gaussian process mixture model-based hard-cut iterative learning algorithm for air quality prediction, Appl. Soft Comput. 85 (2019) 105789.
17. M. Péres, G. Ruiz, S. Nesmachnow, A.C. Olivera, Multiobjective evolutionary optimization of traffic flow and pollution in Montevideo, Uruguay, Appl. Soft Comput. 70 (2018) 472–485.