ISSN: 2454-9940

IJASEM

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

E-Mail :
editor.ijasem@gmail.com
editor@ijasem.org

www.ijasem.org

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

ISSN 2454-9940

www.ijasem.org

Vol 19, Issue 2, 2025

Machine Learning-Based Air Quality Forecasting

[1]Mohammed Khaja Moizuddin, [2] Mohammed Sadath Ali, [3] Mohammed Naif Ullah, [4] Mohammed Asim Ali.
[1]Assistant Professor,Department of CSE-AIML, Lords Institute of Engineering & Technology.
[234] Student Department of CSE-AIML, Lords Institute of Engineering & Technology.

Abstract: There is little question that everyone has an obligation to do all they can to keep the air they breathe clean, and that access to clean air is a fundamental human right that is fundamental to the idea of citizenship.The primary solution to early warning and pollution control has been investigated as air quality prediction. In order to forecast air quality, we provide a method that uses a machine learning framework called the sunlight GBM model in this article. In order to improve the accuracy of air quality predictions, our model—trained using a lightweight GBM classifier—combines meteorological information from many sources. Periodic air quality observance data is used to forecast the future trend of air pollutants using the existing network of air quality observation stations and satellite meteorologic information. With an accuracy rate of 92%, the predicted system was determined to administer

Keywords: **Air Pollution, Decision Tree, Linear Regression, Machine Learning, Random Forest, Supervised Learning, SVM.**

INTRODUCTION

Harmful air pollution is an international epidemic that affects a large number of nations. Most major cities are seeing increased levels of ground-level air pollution as a result of global economic and social growth, particularly in rapidly emerging nations like China and India. While everyone is vulnerable to the health effects of air pollution, those with preexisting conditions like heart disease or lung illness are at a much higher risk. In order to save humans from the harmful effects of air pollution, it is crucial to construct an early warning system that not only gives accurate predictions but also notifies local populations of health alarms. Every year, around 7 million people die too soon as a result of air pollution, which includes both outdoor and indoor sources. The purpose of this study is to develop a dataset for use in making forecasts about future air pollution levels. By working with them, we can get the most accurate predictions by comparing several models and then finding the best answers. Create a reliable software by combining machine learning algorithms with other methods applied to massive datasets; identify the best way to improve air quality for people; and Using methodological factors, it is used to forecast future air pollution concentrations. Most notably, there are oxides (NO), monoxides (CO), particulate matter (PM), sulfur dioxide (SO2), and so on. The incomplete oxidation of propellants (rock oil, gas, etc.) results in the production of carbon monoxide. Ignition of thermal fuel produces nitrogen oxide; carbon monoxide causes headaches and vomiting; smoking produces aromatic hydrocarbons, which disrupt metabolic processes; gas oxides cause vertigo and nausea; and anything with a diameter of 2.5 micrometers or smaller affects human health in additional ways. We need to do something about the air pollution that's already there. The AQI is a measure of the air quality standard. Classical methods, such as probability and statistics, were formerly used to forecast air quality, but they were very laborious. These days, it's quite easy to get data on air pollution from sensors, all thanks to technological advancements. Thorough analysis is required for data assessment in order to identify contaminants. A process called convolution In order to execute the appropriate actions in response to future AQI predictions, algorithms for machine learning, deep learning, algorithmic neural networks, and neural networks guarantee success. Within the realm of computers, there are three types of learning algorithms: supervised learning, unsupervised learning, and reinforcement learning. These algorithms make up machine learning. We have employed the supervised learning method in the proposed work.

III – BACKGROUND A wide range of applications rely on machine learning to discover optimal solutions to real-world problems. Algorithms for machine learning can learn new things without human intervention. The field of machine learning makes use of three distinct kinds of algorithms in a wide range of contexts.

1. An algorithm for Supervised ML
2. Deep Learning Without Human Supervision
3. Machine for Reinforcement Leaning

1. Linear Regression: Predicting actual values from continuous variables is the job of Linear Regression. Numerous fields make use of it, including healthcare, economics, and finance.

Basic Principle of Linear Regression: In order to do a linear regression or discover the link between several independent and dependent variables, four assumptions must be satisfied.

1. The consistency of variation
2. Independence
3. A linear relationship
4. Normality

2. Support Vector Machine:

As an SL algorithm, support vector machine (SVM) draws a line separating the two classes, thus splitting the plane in half. The hyperplane is the line that cuts the plane in half. Distances from data points to separation lines are always given in a perpendicular fashion. Linear and nonlinear classification are both within its capabilities. Its primary function is to do regression and classification.

3. Decision Tree

A supervised learning technique that represents a choice based on a condition is the decision tree. Both regression and classification make use of it. In every case, the decision tree is built from the very top down. A node is considered root node if it is the first node from the top. "Leaf node" describes the very last node. Connected to both the root and leaf nodes are internal nodes. The internal nodes are divided and judgments are made based on certain conditions. In real time, both the tree size and the method complexity rise with an increase in the number of variables. Classification trees and regression trees are the two main varieties of decision trees. In order to facilitate data analysis, a classification tree is used to categorize the dataset. We are unable to make a forecast using this technique, however.

4. Randorm Forest:

Decision Tree In its most basic form, it is a collection of decision trees used for classification and regression. To determine the voting majority, classification is used. The mean value is calculated using regression. Better accuracy, better robustness, and compatibility with different types of data (binary, category, and continuous) are all features of this technique. A random forest is just a collection of decision trees. When training, only 75% of the dataset is taken into account. The training data is randomly sampled, and the Random Forest is used to build several decision trees depending on the attributes sampled.

IV- PROPOSED SYSTEM

Using a single schema, the sensors collect data on air contaminants, which are then stored as a dataset. Various capabilities, such as standardization, attribute choice, and discretization, were used to preprocess this dataset. The dataset is divided into a coaching dataset and a check dataset once it is created. On top of that, the training dataset is used by any supervised machine learning algorithm. After analysis, the acquired findings square measure in agreement with the testing dataset. The suggested model's layout is shown in Fig. 1.

Step 1: Copies of past datasets extracted.

Step 2: Cleaning and preparing data for analysis.

Step 3: In a 70:30 ratio, split the dataset.

Step4: Apply feature selection to the features in the dataset.

Step5: Use several regression methods for training and testing.

1. System Architecture:

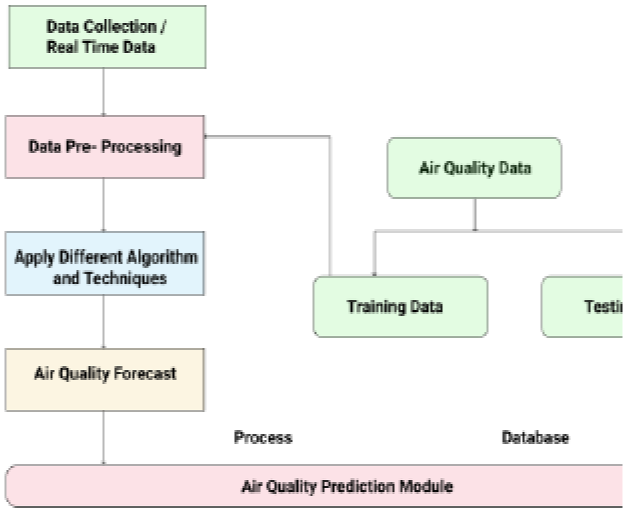

Figure: architecture of Air Quality Prediction

METHODOLOGY

The system consists of two main phases: the first is training, during which the system learns to use the data set to its advantage by fitting a model (line or curve) that is based on the rules that were previously determined. Second, we put the system through its paces by feeding it data and seeing how it responds. The precision is verified. Consequently, the data used to either train the model or verify its acceptability. Since the system's purpose is to detect and forecast AQI levels, suitable algorithms should be used for these two distinct purposes. Various algorithms were evaluated for their accuracy before being chosen for further use.

IMPLEMENTATION

The line equation for SVR is Y = Wx + b, which is the same as for LR. Hyperplane is the term used to describe this straight line in SVR. For the purpose of plotting the boundary line, the data points on each hyperplane that are closest to it are referred to as support vectors. Within a certain range, SVR attempts to find the optimal line fit, where x is the distance between the hyperplane and the boundary line.

Stage1: Information Gathering: In this step, we are compiling a complete inventory of all the factors that influence air pollution. Pollutants may be detected by a multitude of sensors in smart cities.

Stage2 Data cleansing and missing value filling are part of data preprocessing.

Stage3: Feature Selection with GA: Finding the most relevant inputs for the predictive model is the process of feature selection. Using this method, you may find and eliminate characteristics that are redundant, unnecessary, or otherwise do not improve or detract from the prediction model's accuracy.

Stage4: This level involves estimating air pollution using a random forest algorithm using multivariate multi-step time series data. Several trees are used, with each tree trained on a different subset of the time-series data.

Stage5: Forecast: In this case, our method forecasts the level of air pollution.

RESULT & DISCUSSION

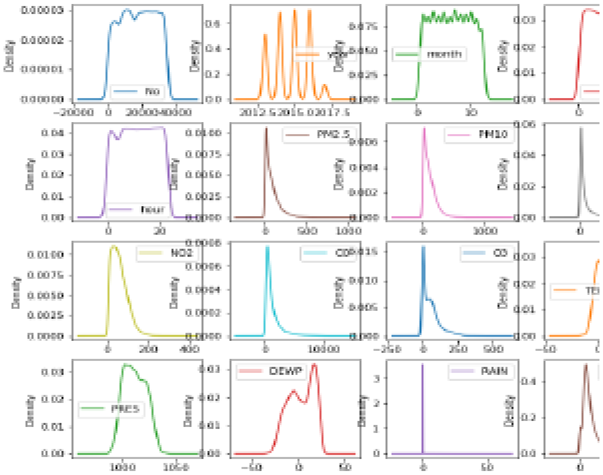

Figure: Pair plots of Air Quality

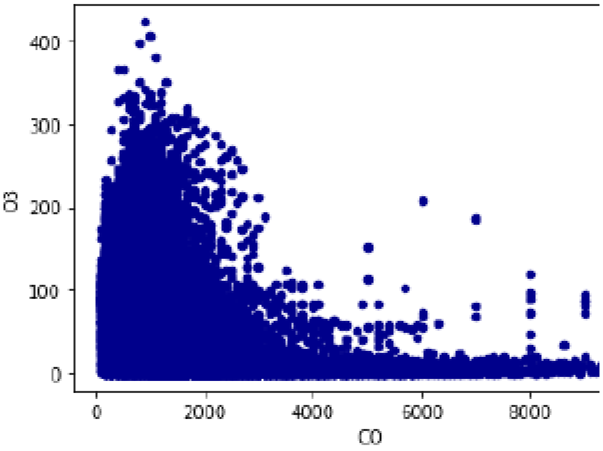

Figure: Air Quality Prediction

V- CONCLUSION

Building a reliable module to forecast air pollution and risk is the primary objective of this project. We take into account the elements that are used for prediction.. In order to make the most accurate predictions possible about the air quality, a prediction model has been developed. Machine learning algorithms and techniques are used to analyze and forecast risk factors and calculate air quality based on a small number of strongly linked characteristics.

VI- BIBLIOGRAPHY

.

[1]Verma, Ishan, Rahul Ahuja, HardikMeisheri, andLipikaDey. ”Air pollutant severity rediction using Bi-directional LSTM Network.” In 2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI), pp. 651-654. IEEE, 2018.

[2] Figures 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.

[3] Yang, Ruijun, Feng Yan, and Nan Zhao. ”Urban air quality based on Bayesian network.” In 2017 IEEE 9th Fig. 10. RH w.r.t NO Fig. 11. RH w.r.t NO2 International Conference on Communication Softwareand Networks (ICCSN), pp. 1003-1006. IEEE,2017.

[4] Ayele, TemeseganWalelign, and RutvikMehta.”Air pollution monitoring and prediction using IoT.” In 2018 Second International Conference on Inventive Communication 6 Fig. 12. RH w.r.t Temperature Fig. 13. RH w.r.t CO and Computational Technologies (ICICCT), pp. 1741-1745. IEEE,2018.

[5] Djebbri, Nadjet, and MouniraRouainia. ”Artificial neural networksbased air pollution monitoring inindustrial sites.” In 2017 International Conference on Engineering and Technology (ICET), pp. 1-5. IEEE,2017. **www.ijcrt.org © 2021 IJCRT | Volume 9, Issue 7 July 2021 | ISSN: 2320-2882 IJCRT2107446 International Journal of Creative Research Thoughts (IJCRT) www.ijcrt.org e213**

[6] Kumar, Dinesh. ”Evolving Differential evolution method with random forest for prediction of Air Pollution.” Procedia computer science 132 (2018): 824-833.

[7] 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.

[8] Svetnik, Vladimir, et al. ”Random forest: a classification and regression tool for compound classification and QSAR modeling.” Journal of chemical information and computer sciences 43.6 (2003): 1947-1958.

[9] Biau, GA˜ Srard. ”Analysis of a random forest model.” ˇJournal of Machine Learning Research 13.Apr (2012): 1063- 1095.

[10] Biau, Gerard, and ErwanScornet. ”A random forest ´ guided tour.” Test 25.2 (2016): 197-227.

[11] Grimm, Rosina, et al. ”Soil organic carbon concentrations and stocks on Barro Colorado Island— Digital soil mapping using Random Forests analysis.” Geoderma 146.1- 2 (2008): 102-113.

[12] Strobl, Carolin, et al. ”Conditional variable importance for random forests.” BMC bioinformatics 9.1 (2008): 307.

[13] Svetnik, Vladimir, et al. ”Random forest: a classification and regression tool for compound classification and QSAR modeling.” Journal of chemical information and computer sciences 43.6 (2003): 1947-1958.

[14] Verikas, Antanas, AdasGelzinis, and MarijaBacauskiene. ”Mining data with random forests: A survey and results of new tests.” Pattern recognition 44.2 (2011): 330-349.

[15] Ramasamy Jayamurugan,1 B. Kumaravel,1 S. Palanivelraja,1 and M.P.Chockalingam2 International Journal of Atmospheric Sciences Volume 2013, Article ID 264046, 7 pages http://dx.doi.org/10.1155/2013/264046