ISSN:0976-0172

Journal of Bioscience And Technology
www.jbstonline.com

Predicting Air Quality with Machine Learning

[1] M. Pradeepthi, [2] P. Madhuri,

[1]Assistant Professor, Megha Institute of Engineering & Technology for Women, Ghatkesar.
[2] MCA Student, Megha Institute of Engineering & Technology for Women, Ghatkesar.

Article Info

Received: 30-04-2025 Revised: 16-06-2025 Accepted: 28-06-2025

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 this research, we provide a machine learning framework called the sunshine GBM model to forecast air quality, which we call an Associate in Nursing air quality prediction system. By making full advantage of accessible abstraction data, our model—trained using a lightweight GBM classifier—increases the prediction accuracy of air quality predictions by combining meteorological information from several 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. An accuracy rate of 92% was observed in the predicted system's administration of nursing.

Keywords:

Subjects: Pollutants in the Air, Decision Tree, Linear Regression, ML, SF, SVM.

I-INTRODUCTION

Numerous nations face the issue of air pollution, which has negative effects on people's health. Greater concentrations of ground-level air pollutants are becoming more common in most major cities as a result of global economic and social growth, particularly in rapidly emerging nations such as 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. An early warning system that not only gives accurate predictions but also notifies local residents of health warnings would offer essential information to safeguard humans from the harm caused by air pollution, which may have more severe repercussions. About 7 million people die too soon each year as a result of indoor and outdoor air pollution. 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. Discover the best solution for air quality that benefits humans by developing a robust application utilizing machine learning algorithms and various methodologies applied to massive datasets. It is used for the purpose of forecasting the upcoming levels of air pollution based on methodological characteristics. Most notably, there are oxides (NO), monoxides (CO), particulate matter (PM), sulfur dioxide (SO2), and so on. When propellants like rock oil, gas, etc., undergo incomplete oxidation, they produce carbon monoxide. Gas oxides produce vertigo and nausea; carbon monoxide causes headaches and vomiting; aromatic hydrocarbons are produced by smoking and cause problems with metabolic processes; and nitrogen oxides are produced when thermal fuel is

ISSN:0976-0172

Journal of Bioscience And Technology
www.jbstonline.com

ignited. Stuff with a diameter of 2.5 micrometers or smaller affects human health adversely. We need to do something about the air pollution that's already there. One way to evaluate the air quality is by looking at the Air Quality Index (AQI). Classical methods, such as probability and statistics, were formerly used to forecast air quality, but they were very laborious. Data about air pollution retrieved from various sensors is now quite easy to get by, all thanks to technological advancements. Thorough data analysis is required for the purpose of identifying the contaminants. Fuzzy logic In order to execute the appropriate actions in response to future AQI predictions, neural networks, algorithmic neural networks, deep learning, and machine learning algorithms guarantee success. In the field of computers known as machine learning, there are three main types of learning algorithms: supervised learning, unsupervised learning, and reinforcement learning. In our planned study, we mostly used the supervised learning method.

II – 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. There are three main categories of machine learning algorithms, each with its own set of uses in the field. one. An algorithm for supervised machine learning 2. Machine Learning Without Experimentation 3. A Machine That Leans In 1. Linear Regression: This method works by taking continuous variables and applying them to make predictions about the actual values. The fields of economics, finance, healthcare, and many more make use of it. Linear Regression Assumption: In order to do linear regression or discover the link between several independent and dependent variables, four assumptions must be met. 1. Variance homogeneity 2. Autonomy 3. Predictability 4. Consistency The second SL technique is Support Vector Machine (SVM), which uses a line to split the plane in half at the class boundaries. The term "hyperplane" describes 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. Classification and regression are its primary applications. 3. A Decision Tree Among the many supervised learning techniques, Decision Tree is a useful tool for visually representing condition-based decisions. Its applications include regression and classification. In every case, the decision tree is built from the very top down. A root node is the very first node in a tree. "Leaf node" describes the very last node. The space between a node's root and its leaf is occupied by an internal node. After dividing the internal nodes according to certain conditions, choices are made. When dealing with a growing number of variables in real time The algorithm becomes more complicated as the tree gets bigger. Classification trees and regression trees are the two main varieties of decision trees. To facilitate data analysis, a classification tree is used to categorize the dataset. However, it is not possible to make a forecast using this method.4) Randorm Forest, also known as Random Forest 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. Using regression, one may derive the average. Better accuracy, better robustness, and compatibility with different types of data (binary, category, and continuous) are all features of this technique. Simply said, Random Forest is an array of decision trees. For training purposes, 75% of the dataset is taken into account. After randomly selecting certain attributes from the training data, the Random Forest is used to build various decision trees.

III- PROPOSED SYSTEM

Data on air pollutants is collected from sensors, processed using a consistent schema, and stored as a dataset. Standardization, attribute choice, and discretization are some of the distinct preprocessing features used to this dataset. A coaching dataset and a check dataset are created from the prepared dataset. Additionally, the training dataset was subjected to any supervised machine learning algorithms. After analysis, the acquired findings square measure in agreement with the testing dataset. You can see the suggested model's layout in Figure 1. First step: Gathering historical data. Data pre-processing and normalization is the second step. Step 3: Assign a ratio of 70:30 to the dataset. Step4: Select characteristics from the dataset. Step 5: Experiment with various regression methods for training and testing.

System Architecture:

ISSN:0976-0172

Journal of Bioscience And Technology
www.jbstonline.com

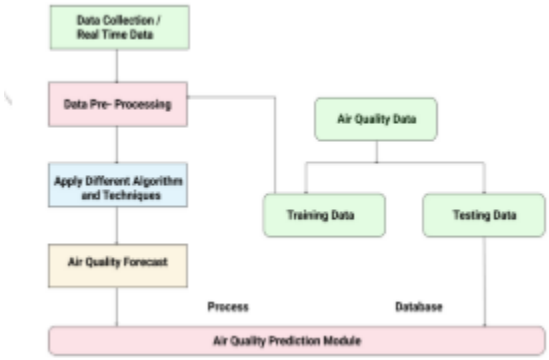

Figure: architecture of Air Quality Prediction

IV. METHODOLOGY

The system consists of two main phases: one. During the training phase, the system is taught to use the data set by analyzing it and then fitting a model (line or curve) that is based on the rules that were selected. Second, we put the system through its paces by feeding it data and seeing how it responds. The precision is verified. Therefore, the data used to either train the model or verify its suitableness. Since the system's purpose is to detect and forecast AQI levels, suitable algorithms should be used for these two distinct purposes. An evaluation of the algorithms' accuracy was conducted prior to their selection for further usage.

V. 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. Using a threshold value, which is the distance between the hyperplane and boundary line, SVR attempts to fit the optimal line. The first step is to gather all the data on the factors that contribute to air pollution. In smart cities, there are a lot of sensors that can detect pollution. In the second stage, data preprocessing, missing values are filled in and noise is removed from the data. Third Stage: GA-Based Feature Selection: In order to discover the most useful inputs for the prediction model, feature selection must be performed. If you want to make sure your predictive model is as accurate as possible, you may use this method to find and delete characteristics that aren't necessary, irrelevant, or redundant. Fourth Stage: Random Forest-Based Multivariate Multistep Time Series Prediction: At now, we are forecasting air pollution using multivariate multi-step time series data and a random forest method. Each of the many trees has its own set of time-series data from which it was trained. Phase 5: Forecasting This is where our algorithm makes air pollution predictions.

VI. RESULT & DISCUSSION

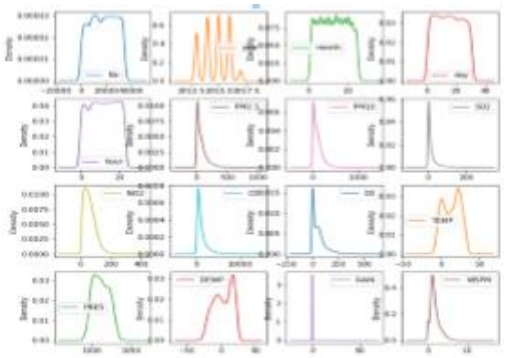

Figure: Pair plots of Air Quality

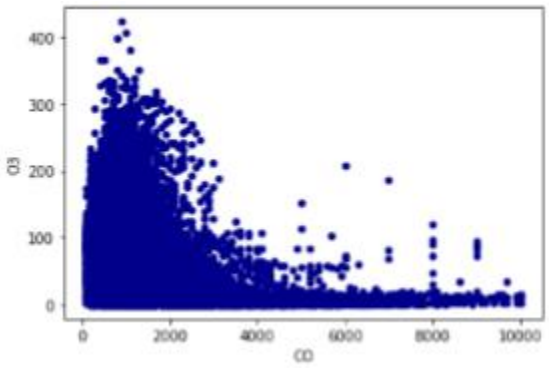

Figure: Air Quality Prediction

VII- CONCLUSION

The overarching goal of this project is to build a reliable module that can forecast air pollution and risk. We take into account the characteristics that are used for prediction. In order to make the most accurate predictions possible about the air quality, a prediction model has been developed. In order to analyze and anticipate risk factors and calculate air quality using machine learning algorithms and techniques, just a small number of strongly linked characteristics are used.

VIII- BIBLIOGRAPHY

ISSN:0976-0172

Journal of Bioscience And Technology
www.jbstonline.com

[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.

[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): 102113.

[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