International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 2 Volume No.7 March - April – 2023
DOI:10.46647/ijetms.2023.v07i02.091 ISSN: 2581-4621

AIR POLLUTION PREDICTION USING MACHINE LEARNING

Sneha A R[1], Premkumar M[2] , Sanjuvikashini A P[3] ,Soniga N[4], Swathika T[5]
[1,3,4,5]UG – Computer Science and Engineering, KPR Institute of Engineering and Technology, Arasur, Coimbatore
[2]Assistant Professor, Artificial Intelligence and Data Science, KPR Institute of Engineering and Technology, Arasur, Coimbatore

ABSTRACT
The fact that environment tracking is focused largely on the fundamental rights of people, lifestyles, and health makes it so important. As a result, this device tracks the quality of the air using excellent sensor nodes within that check for CO2, NOx, UV light, temperature, and humidity. The gad get is able to categorize automatically if a certain geographic area is going above the established gas emission restrictions thanks to the statistics assessment using device mastering algorithms. In order to choose the most contaminated sectors, the DB SCAN with LR, SVM, and NB set of rules delivered a noteworthy category overall performance. Monitoring air quality is a crucial concern in many commercial and physical areas of the world. In areas with serious difficulties with air pollution, Air Quality Operational Centers (AQOCs) are established specifically for this purpose. The AQOCs are operational units responsible for managing tracking networks, analyzing the gathered data, and eventually disseminating online assessments of air pollutants and their short- and long-term evolution. Up until recently, modelling of air pollution events has been focused mostly on dispersion models, which approximate the complex physicochemical processes at play. Although the intricacy and complexity of these models have increased over time, their application in real-time atmospheric pollution tracking appears to no longer be acceptable in terms of performance, input data requirements, and compliance with the problem's time limitations.

1. Introduction
Many nations across the world are affected by air pollution, which can have deadly consequences for human health. Over the past century, our atmosphere has gotten worse as a result of an increased reliance on fossil fuels. A number of vehicle types have a significant impact on pollution. The primary contributors to air pollution, which can affect people's health in the short- and long-term, are RSPM, SO2, NO2, SPM, and other pollutants. The objective of this study is to ascertain if analytics methods can be utilized to create a system that can approximatively predict future pollution levels with a high degree of accuracy. It is shown that strategies for rendered linear regression are insufficient for the time- dependent data. With a high degree of confidence, we have predicted future levels of several pollutants using a time series forecasting method. The efficiency of our recommended method employing SARIMA is demonstrated by the experimental examination of the forecasting for the levels of air pollution in Bhubaneswar City. to manually create the necessary algorithms for more complicated tasks. In actual use, it could be more effective to assist the computer in developing its own algorithm as opposed to having human programmers specify every essential step

2. Machine Learning
Science's field of machine learning enables computers to learn without explicit programming. One of the most intriguing new technologies is machine learning. The computer's ability to learn, as the name suggests, is what gives it a more human-like personality. Machine learning is now being actively used in more places than one may think. Machine learning programmers are capable of doing tasks without having them explicitly written. For certain tasks, computers use the data that is readily available to learn. For simple tasks given to computers, it is possible to create algorithms that tell the machine exactly what to do in order to solve

DOI:10.46647/ijetms.2023.v07i02.091 ISSN: 2581-4621

3. SARIMA

Science's field of machine learning enables computers to learn without explicit programming. One of the most intriguing new technologies is machine learning. The computer's ability to learn, as the name suggests, is what gives it a more human- like personality. Machine learning is now being actively used in more places than one may think. Machine learning programmes are capable of doing tasks without having them explicitly written. For certain tasks, computers use the data that is readily available to learn. For simple tasks given to computers, it is possible to create algorithms that tell the machine exactly what to do in order to solve the problem at hand; the computer doesn't need to learn anything. A human could find it challenging to manually create the necessary algorithms for more complicated tasks. In actual use, it could be more effective to assist the computer in developing its own algorithm as opposed to having human programmers specify every essential step. the problem at hand; the computer doesn't need to learn anything. A human could find it challenging

4. Related Work

4.1 Air quality evaluation and pollutant concentration prediction in a new air quality monitoring and early warning system.

Wang Jian - An air quality system that includes both air quality assessment and air pollutant concentration predictions is uncommon, nevertheless. Within this piece, they suggest the novel in the air surveillance and effective screening method. Using vague thorough processes, the main contaminants are recognized and the level of air pollution is more thoroughly assessed. An Elman neural network, a modified cuckoo search method, and differential evolution were used to boost the precision of six main air systems. An original hybridization model with complementing ensemble empirical mode decomposition, pollutant concentrations, and all of these are described. To validate the effectiveness of the approach, pollution statistics from two Chinese cities are used. The fuzzy comprehensive evaluation's findings indicate that PM10 and PM2.5 are the main contaminants in the Xi'an and Jinan air respectively, consequently that Xi'an has superior air quality than Jinan. A Foresight outcomes show which is suggested inverters, with its improved prediction accuracy and stability, is unquestionably better than all comparison models.

The Chinese people wish to clear the air and get rid of the fog. More than 2700 monitoring stations, more than 268,000 pieces of monitoring equipment, and more than 60,000 monitoring staff have all been established to achieve this (MEP Ministry of Environmental Protection, 2015). China has made great strides toward clearing smog and enhancing air quality, but, the continued difficulties of reducing air pollution still require attention. Recent years have seen a significant increase in research on air quality monitoring and evaluation systems. Forecasting air pollutants and evaluating air quality can be done separately in this study.

4.2 A Case Study Applied to Tehran, Iran's Capital: A Novel Approach for Improving Air Pollution Prediction Based on Machine Learning Approaches.

Abdullah Mohammed Delavar - A few of the primary issues in urban cities all across the world has been highlighted as air pollution, particularly in Tehran, Iran’s capital, which inhabitants' health difficulties and other air pollution-related harm have long been a source of contention for the city's government and residents. The study area indicates that PM10 and PM2.5 particles are a substantial contributor to the air pollution in Tehran. The present study’s objective was to create prediction models for calculating Tehran's air pollution levels based on PM10 and PM2.5 pollution concentrations. The input parameters included information on the day of the week, the month of the year, the terrain, the weather, and the amount of pollution in the two closest neighbours. In order to predict air pollution, machine learning techniques such as regression support vector machines, spatially weighted regression, artificial neural networks, and auto-regressive nonlinear neural networks with an external input have been used. The error percentagewas later decreased and improved by 57%, 47%, 47%, and 94%, respectively, by using a prediction model with the previously mentioned methodologies. The suggested autoregressive nonlinear neural network with external input is the most accurate method for predicting air pollution, with a one-day prediction error of 1.79

DOI:10.46647/ijetms.2023.v07i02.091 ISSN: 2581-4621

g/m3.The evolutionary algorithm was used to identify the most reliable variables for predicting air pollution, whichincluded information on the day of the week, month of the year, terrain, wind direction, maximum temperature, and pollutant rate of the two closest neighbors. Municipal administrators can utilise the pollution statistics to notify the public about the issue of air pollution by including the density of daily PM2.5 and PM10 pollutants. By using public transit and avoiding polluted areas, people may be able to reduce the amount of pollution. PM10 and PM2.5 make up the majority of the air pollution concentration in Tehran, according to the most recent data available from 21 stations maintained by the Tehran Air Quality Control Company(AQCC) and 16 sites regulated by the Iranian production company for the environment. The most common pollutant is PM2.5, which is then followed by CO, O3, NO2, SO2, PM10, and PM2.5. Nearly 5% of PM2.5 pollutants come from surrounding inhabited regions in the west,notably the cities of Karaj, Shahryar, and Rey, according to AQCC study done in 2017 and a technical report on the Tehran air pollution forecast system was produced. According to the AQCC's expert opinion, when compared to other pollutants that have been found to be under 5%, the amount of PM2.5 pollution shown here is the highest.

4.3 Air Pollution Monitoring and Forecasting System Based on IOT.

Chen Xiaojun- According to empirical study, typicalair automated monitoring systems offer high precision, but because of their bulky design, high cost, and single datum class, they are unsuitable for large-scale installation. This study suggests a method for monitoring and forecasting air pollutionin real-time that is based on the pertainingto the utilization of Internet of Things (IOT) for environmental protection. By utilizing IOT, the hardware cost of this system may be reduced by half.A network of monitoring sensors may be built using the technology in a large-scale monitoring area. In addition to performing the functions of a standard automated air monitoring system, it also demonstrates the capacity to forecast the evolution of air pollution over a specified time period by using neural network technology to analyze the information gathered by a front-end perception system. Targeted emergency disposal techniquesmay be used to limit losses during real usage. A laboratory analysis reveals that the conventionalair automated monitoring system consists of pricey, large, and high-tech equipment. Due to thehigh cost and size, installation on a large scale is not viable. This technology can only be installed in
crucialmonitoring locations for a small number of crucial industries, hence there are no system data accessible to predict the overall technology provides for the flexible and pollution picture.

This paper presents a strategy that addresses shortcomings in traditional monitoring systems and detection techniques while reducing test costs. It blends IOT technology with environment monitoring. IOC cost-effective installation of sensors across the region for omnidirectional monitoring and data support by replacing the standard empirical analysis monitoring equipment with sensor networks.

4.4 New Multilinear Regression Equations for Lateral Spread Displacement Prediction.

Leslie Bartlett, and Youd developed the first empirical formulations for lateral spread displacement prediction in 1992 and 1995; they are now frequently utilized in engineering practise. Multilinear regression (MLR) was used to generate the equations using a big case history database. Thisreport provides an update and correction to the first analysis. Here, the errors are cleared up. By mistake, Bartlett and Youd overstated the reported displacements for the lateral spreading caused by the 1983 Nihonkai- Chubu, Japan earthquake. When boundary shear prevented free lateral movement, some spots were removed. It was uploaded the data from three more earthquakes. In order to avoid unnecessarily overestimating displacements as R goes smaller, the mean-grain- size term's functional form has been changed to depend on the earthquake's magnitude.Regressing the modified data resulted in new MLRequations. The use of the new equations is advised for engineering applications. An empirical equationfor forecasting lateral spread displacement at liquefiable locations was developed by Bartlett and you in the early 1990s.

Multilinear regression was used to generate theequation using a substantial case history database that theseresearchers had obtained. Over time, we have become aware of certain improvements and alterations to the equation that are required. These alterations and corrections have been made.

International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 2 Volume No.7 March - April – 2023
DOI:10.46647/ijetms.2023.v07i02.091 ISSN: 2581-4621

4.5 Categorization, fuzzy sets, and multilayerperceptron's.

Mitra S - This paper presents a multilayer perceptron- based fuzzy neural network model that can categorize patterns in a fuzzy way and uses the backpropagation approach. The input vectorconsists of membership values for linguistic attributes, whereas the output vector is described in terms of fuzzy class membership values. Consequently, by properly weighing the back propagated errors basedon the membership metrics at the different outputs; fuzzy uncertain patterns may be properly described. The learning rate is gradually decreased over the course of training in discrete steps until the network converges to a minimum error solution. The outcomes are contrasted with the Bayes classifiers, traditional MLPs, and outcomes from other comparable models.

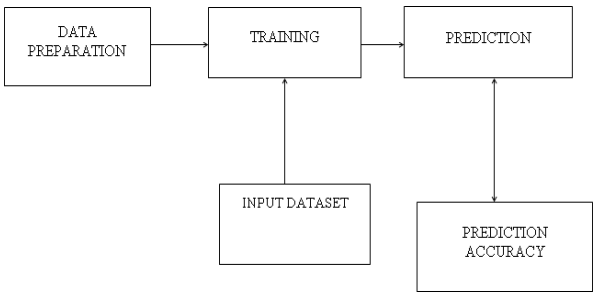
5. Existing System

Choosing the system's software and hardware architecture, as well as which functionalities should be implemented in software running on programmable components and which should be done in more specialized hardware, is an essential part of the design problem. Embedded systems are typically employed in circumstances where dependability and safety are more important than performance. Today, manual design and prior experience with goods that are similar to them have a significant impact on the on-the-fly creationof embedded systems.

6. Proposed System

Effectively, however in our project, we used a convex hull to generate the statistics shape, allowing us to use a set of statistics that included date, time, temperature, CO2, NO2, O3, and PM10 in addition to SO2. With this method, we were able to create a structural database out of a sizable dataset by combining DBSCAN with LR, SVM, and NB to obtain statistics on the structuralarrangement of air molecules. Now that DBSCAN can supply a specified value, we are able to form clusters. (DBSCAN) is a collection of suggested statistical clustering rules. It is a density-based, fully clustering, non-parametric set of rules: given afixed number of factors in a particular area, it groups together factors that are tightly packed together (factors with many near neighbours), designating as outlier's factors those that locate solely in low-density areas. One of the most popular common place clustering methods and one that is also frequently cited in medical literature is DBSCAN. After the system is trained, the input data are analysed using a variety of machine learning methods to give accurate findings. We utilise the Indian dataset, which consists of diverse data about air quality.

7. Block Diagram



8. Implementation Data

The records had been received from sources. One carries Air fine records and the opposite carries Meteorological Data. Air fine records: From the website Air Now, we collected the historical air fine data for New Delhi. This dataset includes year, month, day, hour, AQI cost for every three hours beginning at three am on 1-1-2015 to 24-+4-2017, among other columns, as well as functions for our use. There are 6700 values total in this. Some columns, like conc and gadgets, etc., might not be necessary for us. For our painting requirements, we just augment the meteorological records by adding the values from the AQI column for the hour, month, and day.

International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 2 Volume No.7 March - April – 2023
DOI:10.46647/ijetms.2023.v07i02.091 ISSN: 2581-4621

9. Meteorological Data

The weather information for Delhi was updatedand taken from Kaggle. It includes Delhi's hourly weather information from 1997 to 2017. From January 1st, 2015, through April 24th, 2017, we reduced it to every three hours of data. Date, time, cords, dewpt, fog, hail, heatindexm, hum, precept, pressure, rain, snow, tempm, thunder, tornado, vims, odd, wind gust, wind-chill, and wisdom are the columns in the climatic facts. Conditions such as fog, partly fog, mist, haze, light drizzle, rain, etc. are described by Cond's. Dew point is provided by dewpt, and fog and hail are also indicated by the presence or absence of fog. Precept describes the precipitation; hum describes the humidity. Pressure is used to describe pressure. The direction of the wind is represented by odd, and its speed by wisdom.

10. Data Visualization

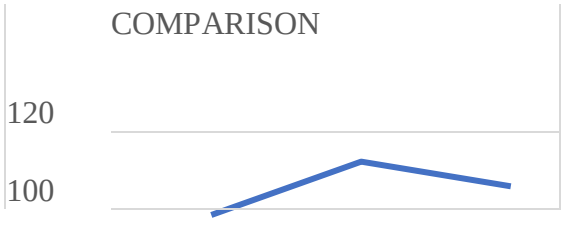
As their AQI values were no longer available, we first eliminate a few initial entries from the dataset. The data is then visualized to help us understand how different capabilities affect the AQI tiers over time. We want to identify the desired and unnecessary skills for our task when we plot the characteristic versus AQI graph. The plots of the attributes that seem to affect AQI values.

11. Data Cleaning

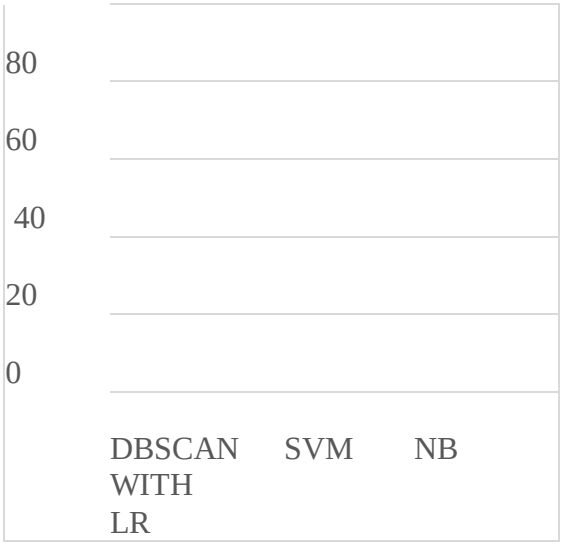
We can see from the plots that our desired capabilities are date and time, conditions, dewptm, hum, pressure, temperature, wdird, wspdm, month, day, and hour, and AQI. The AQI clearly decreases when the windspeed is better. A reliable forecastingis probably wind speed. Similar to how AQI scores are a little better in the winter, air flow is important as well. Every other capacity frequently performs below averagely or no longer significantly affects the AQI rate. The ones columns are therefore removed. We then use transferring periods to establishprevious fee capabilities, in order thatprimarily based totally on preceding five givenclimate and AQI records. We can anticipatethe AQI rate for the current hour. Then, if either removal of those rows or filling them with the values from the imply column proven to be more successful, we deal with missing values or NaNs. These actions prepare our records for incorporation into the models.

12. Result Analysis

We use conventional plotting tools and Sclera packages to execute it on an x64 system running Python 2.7, with 8GB of RAM, and an Intel i7processor. Using the Panda's data frame, the major part of the task was completed. With a ratio ofaround 7:3, the data was divided into two sets: a train set and a test set. . The test set, which contained 4400 train values and 2000 test values, was used to evaluate each model after it had been trained using the train set. The graphs provide the regression plots for each of the models we used to forecast the AQI. Additionally, the feature significance analysis has made use of the decision forest and Extra trees. Theresults' accuracy scores and RMSE values are also included. We can see from the regression graphs that the models produce quite accurate results. The blue line represents the corresponding predicted values in ascending order, and the orange line represents the test set points (set in ascending order), according to DBSCAN with Linear Regression and SVM (support vector Machine) and(Navie Bayes) NB. These methods each have adifferent level of accuracy for predicting air quality, with SVM (support vector Machine) offering the highest accuracy values. The features are listed by the Extra Trees model in decreasing order of relevance as follows: month, temperature, conditions, hour, pressure, and humidity readings from the past.



International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 2 Volume No.7 March - April – 2023
DOI:10.46647/ijetms.2023.v07i02.091 ISSN: 2581-4621



Comparison Graph

ALGORITHM	COMPARISON
DBSCAN WITH LR	75.5
SVM	96.75
NB	86.92

Comparison analysis table

CONCLUSION

In this study, we tested the ability of the sklearn library's pre-existing regression models to forecast the values of the air quality index using historical meteorological data. Additionally, we made an effort to identify the traits that some of these models' predictions might benefit the most from. The accuracy of the various machine learning algorithms for the Indian dataset is supplied by the DBSCAN, which offers accuracy of 75%, NB, which offers accuracy of 86%, and the support vector machine, which offers the greatest accuracy of roughly 96%. Additionally, additional data may be used, and real-time prediction techniques can be implemented in settings like Azure ML.

References

1. Environmental research, vol. 158, 2017, pp. 105–117. Air quality evaluation and air pollutant concentration prediction: a unique air quality monitoring and early warning system Yang, Z., andWang, J.

2. Vol. 8, No. 2, page 99, ISPRS InternationalJournal of Geoformation, 2019. An application to Tehran's capital city of an unique technology for improving air pollution prediction based on machine learning methodologies. Along with M. R.Delavar, A. Gholami, and G. R. Shiran, the authors are Y. Rashidi, G. R. Nakhaeizadeh, K. Fedra, and S. HatefiAfshar.

3. IoT-based Air Pollution Monitoring and Forecasting System, 2015 International Conferenceon Computer and Computational Sciences, C. Xiao Jun, L. Xian Peng, and X. Peng (ICCCS). 2015, IEEE, pp. 257–260

4. T. L. Youd, C. M. Hansen, and S. F. Bartlett claim that changes have been made to multi linear regression equations for forecasting lateral spread displacement in Journal of Geotechnical and Geo environmental Engineering, vol. 128, no. 12, 2002, pp. 1007-1017.

5. In 1992, a study titled "Multilayer Perceptron, Fuzzy Sets, Classification" was published by S. K. Pal and S. Mitra.