IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2347-6710

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 13, Issue 4, April 2024

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.423

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 4, April 2024 ||

| DOI:10.15680/IJIRSET.2024.1304051 |

Air Pollution Prediction Using Data Mining

Sanskruti Nichit[1], Gaurangi Mahalle[2], Khushi Thakare[3], Kalyani Sonone[4], Prof. A. Chokhat[5]

Students, Department of Computer Science and Engineering, P. R. Pote (Patil) College of Engineering & Management, Amravati, India[1,2,3,4]

Department of Computer Science and Engineering, P. R. Pote (Patil) College of Engineering & Management, Amravati, India[5]

ABSTRACT: In recent times, due to the vigorous development of industrialization, environmental protection measures can't be effectively guaranteed. Increasingly serious environmental problems have gradually become the primary problem affecting the quality of public life. therefore, we need to establish a fairly accurate air quality prophecy model to understand the possible air pollution process in advance. According to the prophecy results of the model, it's of great significance to establish and take corresponding control measures to reduce air pollution.

In multitudinous artificial and communal regions moment, assessing the air quality has become one of the most important exertions for the dwellers. The multitudinous types of pollution brought on by the use of power, transportation, energy, etc. harm the quality of the air. The accumulation of dangerous feasts is seriously affecting the quality of life in smart cities.

We will bat the Random timber fashion in this suggested system as a means of predicting and measuring pollution in large cities. The data generated by this fashion includes real-time business data, downfall information, and road data. also employed for data training and prophecy is this system. Our proposed machine knowledge-predicated model for predicting air quality had a training delicacy of 99 and a test delicacy of 90. As a result, the created model can produce better air pollution attention auguring performance.

KEYWORDS: air quality prophecy, SOM neural network, NSGAII optimized neural network, Random Forest Algorithm

I. INTRODUCTION

Air pollution is one of the major hazards of environmental pollution. Each living organism needs fresh and good quality air for every second. None of the living goods can survive without analogous air. But because of buses, agricultural exertion, factories and industriousness, mining exertion, and the burning of reactionary powers, our air is getting defiled.

These conditions spread sulfur dioxide, nitrogen dioxide, carbon monoxide, and particulate matter pollutants in our air which is dangerous to all living organisms. The air we breathe every moment causes several health issues. So we need a good system that predicts analogous pollution and is helpful in a better terrain.

This leads us to look for advanced ways of predicting air pollution. So also, we are predicting air pollution for our smart municipality using data mining ways. In our model, we are using a multivariate multistage Time Series data mining fashion using an arbitrary timber algorithm.

Our system takes formerly and current data and applies them to our model to prognosticate air pollution. This model reduces the complexity improves the effectiveness and practicability and can give further reliable and accurate opinions for environmental protection departments for smart cities.

firstly, the SOM neural network model is used for unsupervised clustering of applicable adulterant data to anatomize the correlation between various covered pollutants. Aiming at the problems of a large amount of data and the long calculation time of the algorithm, combined with the clustering results, an NSGA-II optimized neural network is proposed to predict the future pollution situation.

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 4, April 2024 ||

| DOI:10.15680/IJIRSET.2024.1304051 |

II. OBJECTIVES

1) Networks are to record the attention situations of atmospheric pollutants.
2) Locating contamination problem areas and understanding their space-time changes.
3) Complying with atmospheric air protection legislation.
4) estimate the effectiveness of emigration control strategies.
5) Informing citizens regarding the original air quality status.

III. LITERATURE SURVEY

1) Forecasting of daily air quality index in Delhi

AUTHORS Kumar, Anikender, andP. Goyal

As the impact of air adulterants on mortal health through ambient air has attracted important attention in recent times, air quality soothsaying in terms of air pollution parameters has become an important content in environmental wisdom. The Air Quality Index(AQI) can be estimated through a formula, grounded on a comprehensive assessment of the attention of air adulterants, which can be used by government agencies to characterize the status of air quality at a given position. The present study aims to develop a soothsaying model for prognosticating diurnal AQI, which can be used as a base for decision-making processes. originally, the AQI was estimated through a system used by the US Environmental Protection Agency(USEPA) for different criteria adulterants similar to Respirable Suspended Particulate Matter(RSPM), Sulfur dioxide(SO2), Nitrogen dioxide(NO2) and Suspended Particulate Matter(SPM). still, the sub-index and breakpoint attention in the formula are made according to the Indian National Ambient Air Quality Standard. Secondly, the diurnal AQI for each season is read through three statistical models videlicet time series bus-accumulative integrated moving normal(ARIMA)(model 1), top element retrogression(PCR)(model 2), and a combination of both(model 3) in Delhi. The performance of all three models is estimated with the help of observed attention of adulterants, which reflects that model 3 agrees well with observed values, as compared to the values of model 1 and model 2. The same is supported by the statistical parameters also. The significance of meteorological parameters of model 3 has been assessed through top element analysis(PCA), which indicates that diurnal downfall, station position pressure, diurnal mean temperature, and wind direction indicator are maximum explained in summer, thunderstorm,post-monsoon, and downtime independently. Further, the variation of AQI during the weekends(leaves) and weekdays is negligible. thus all the days of the week are reckoned the same in the models.

2) Linear and nonlinear modeling approaches for civic air quality prediction.

AUTHORS Singh Kunwar P., et al

In this study, direct and nonlinear modeling was performed to prognosticate the civic air quality of the Lucknow megacity(India). Partial least places retrogression(PLSR), multivariate polynomial retrogression(MPR), and artificial neural network(ANN) approach-grounded models were constructed to prognosticate the respirable suspended particulate matter(RSPM), SO2, and NO2 in the air using the meteorological(air temperature, relative moisture, wind speed) and air quality monitoring data(SPM, NO2, SO2) of five times(2005 – 2009). Three different ANN models, viz. multilayer perceptron network(MLPN), radial-base function network(RBFN), and generalized retrogression neural network(GRNN) were developed. All five different models were compared for their conception and vaticination capacities using statistical criteria parameters, viz. correlation measure(R), standard error of vaticination(SEP), mean absolute error(MAE), root mean squared error(RMSE), bias, delicacy factor(Af), and Nash – Sutcliffe measure of effectiveness(Ef). Nonlinear models(MPR, ANNs) performed fairly better than the direct PLSR models, whereas, the performance of the ANN models was better than the low-order nonlinear MPR models. Although the performance of all three ANN models was similar, the GRNN outperformed the other two variants. The optimal GRNN models for RSPM, NO2, and SO2 yielded high correlation(between measured and model prognosticated values) of0.933,0.893, and0.885;0.833,0.602, and0.596; and0.932,0.768 and0.729, independently for the training, confirmation, and test sets. The perceptivity analysis performed to estimate the significance of the input variables in optimal GRNN revealed that SO2 was the utmost impacting parameter in the RSPM model, whereas, SPM was the most important input variable in the other two models. The ANN models may be useful tools in air quality prognostications.

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 4, April 2024 ||

| DOI:10.15680/IJIRSET.2024.1304051 |

3) Air pollution modeling for an industrial complex and model performance evaluation

AUTHORS Sivacoumar R, et al,

Jamshedpur, the sword megacity of India positioned in the eastern part of India is affected by adding air pollution situations as a result of concentrated artificial conditioning. The impact of NOX emigrations performing from colorful air pollution sources, viz. diligence, vehicles, and domestic, was estimated using the Industrial Source Complex Short-Term Gaussian dissipation model. The donation of NOx attention from artificial, vehicular, and domestic sources was set up to be 53, 40, and 7. Further statistical analysis was carried out to estimate the model performance by comparing measured and prognosticated NOx attention. The model performance was set up well with a delicacy of about 68.

4) Performance evaluation of air quality models for predicting PM10 and PM2.5 concentrations at civic traffic intersections during the winter period

AUTHORS Gokhale Sharad and Namita Raokhande

Several models can be used to estimate roadside air quality. The comparison of the functional performance of different models' materials to original conditions is desirable so that the model that performs stylishly can be linked. Three air quality models, videlicet the' modified General Finite Line Source Model'(M- GFLSM) of particulates, the' California Line Source'(CALINE3) model, and the' California Line Source for Queuing & Hot Spot computations'(CAL3QHC) model have been linked for assessing the air quality at one of the busiest traffic intersections in the megacity of Guwahati. These models have been estimated statistically with the vehicle-deduced airborne particulate mass emigrations in two sizes, i.e. PM10 and PM2.5, the prevailing meteorology, and the temporal distribution of the measured daily average PM10 and PM2.5 concentrations in wintertime.

The study has shown that the CAL3QHC model would make better predictions compared to other models for varied meteorology and traffic conditions. The detailed study reveals that the agreements between the measured and the modeled PM10 and PM2.5 concentrations have been reasonably good for CALINE3 and CAL3QHC models. Further detailed analysis shows that the CAL3QHC model performed well compared to the CALINE3. The monthly performance measures have also led to similar results. These two models have also outperformed for a class of wind speed velocities except for low winds (< 1 m s− 1), for which, the M-GFLSM model has shown the tendency of better performance for PM10. Nevertheless, the CAL3QHC model has outperformed for both the particulate sizes and for all the wind classes, which therefore can be optional for air quality assessment at urban traffic intersections.

IV. METHODOLOGY

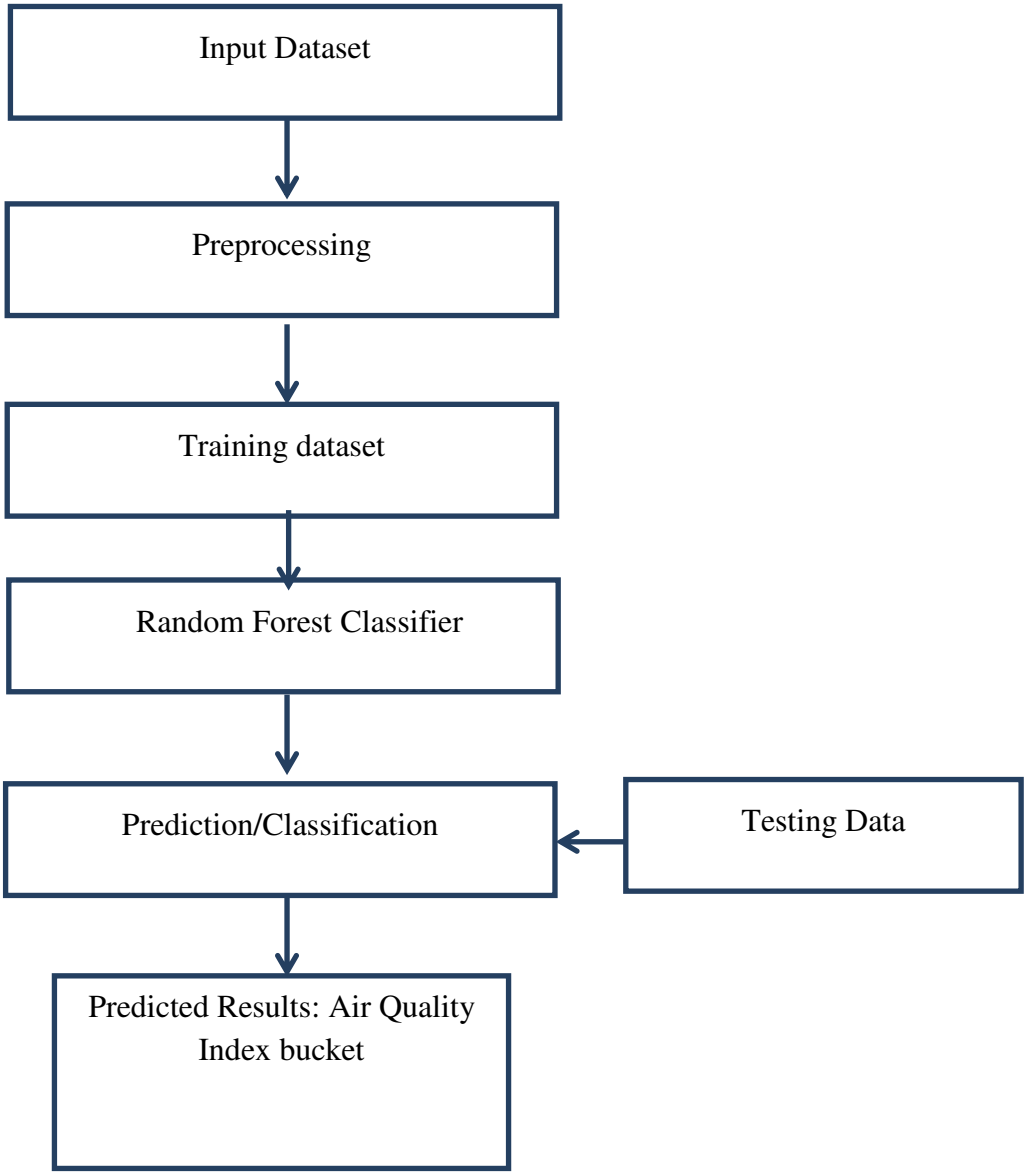
1. Data Collection: Gather air quality data, including variables like meteorological data, emissions data, and historical pollution levels. Ensure data quality and consistency.
2. Data Preprocessing: Clean and preprocess the data, handling missing values, outliers, and data normalization.
3. Feature Selection/Extraction: Choose relevant features or extract new ones that can improve prediction accuracy.
4. Algorithm Selection: Select appropriate data mining algorithms (e.g., decision trees, neural networks, regression models) based on the nature of your data and research objectives.
5. Model Training: Train the selected models using a portion of your data, and fine-tune hyperparameters.
6. Model Evaluation: Assess model performance using metrics like Mean Absolute Error (MAE), Root Mean Square Error (RMSE), or R-squared (R2). Use cross-validation to ensure robustness.
7. Model Validation: Validate the model's performance on a separate dataset to confirm its effectiveness.
8. Prediction and Visualization: Use the trained model to make air pollution predictions, and visualize results through graphs or maps.
9. Interpretability: Explain how the model arrives at predictions, especially if it's a complex machine learning model.
10. Deployment: If applicable, deploy the model for real-time or near-real-time predictions.
11. Documentation: Document your methodology, findings, and code for transparency and reproducibility.
12. Conclusion: Summarize your results, discuss their implications, and suggest potential future research directions.

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 4, April 2024 ||

| DOI:10.15680/IJIRSET.2024.1304051 |

V. DATA FLOW DIAGRAM



VI. BENEFITS

1. It reduces overfitting in decision trees and helps to improve the accuracy
2. It is flexible to both classification and regression problems
3. It works well with both categorical and continuous values
4. It automates missing values present in the data
5. Normalizing of data is not required as it uses a rule-based approach.
6. A random forest produces good predictions that can be understood easily.
7. It can handle large datasets efficiently.
8. The random forest algorithm provides a higher level of accuracy in predicting outcomes than the decision tree algorithm.

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 4, April 2024 ||

| DOI:10.15680/IJIRSET.2024.1304051 |

VII. CONCLUSION

The main advantage of the model is that it can predict the data under different constraints according to the existing data, and can know the possible air pollution process in advance and take corresponding control measures, to improve the ambient air quality. The air quality prediction model proposed and implemented in this paper can greatly improve the prediction accuracy and provide a reference for future research on air quality prediction. The Random Forest Classifier has proven to be a promising method for air quality prediction, producing accurate and reliable results. The algorithm's ability to handle a large number of input variables and its robustness to overfitting make it a suitable choice. For this task, additionally, its interpretability and versatility make it suitable for implementation in a wide range of real-world applications. However, it is important to note that the performance of the Random Forest Classifier can be affected by factors such as the quality and quantity of training data, the choice of hyperparameters, and the specific problem being addressed. Further research and optimization may be needed to fully realize the potential of this method for air quality prediction.

REFERENCES

[1] Carbajal-Hernández, José Juan "Assessment and prediction of air quality using fuzzy logic and autoregressive models." Atmospheric Environment 60 (2012): 37-50.
[2] Kumar, Anikender and P. Goyal, "Forecasting of daily air quality index in Delhi", Science of the Total Environment 409, no. 24(2011): 5517-5523.
[3] Singh Kunwar P., et al. "Linear and nonlinear modeling approaches for urban air quality prediction, " Science of the Total Environment 426(2012):244-255
[4] Sivakumar R, et al, " Air pollution modeling for an industrial complex and model performance evaluation ", Environmental Pollution 111.3 (2001): 471-477
[5] Gokhalesharad and NamitaRaokhande, "Performance evaluation of air quality models for predicting PM10 and PM2.5 concentrations at urban traffic intersection during winter period", Science of the total environment 394.1(2008): 9-24.
[6] Bhanarkar, A. D., et al, "Assessment of the contribution of SO2 and NO2 from different sources in Jamshedpur region, India, "Atmospheric Environment 39.40(2005):7745-India." Atmospheric Environment 39.40 (2005): 7745-7760.
[7] Bhanarkar, A. D., et al, "Assessment of the contribution of SO2 and NO2 from different sources in Jamshedpur region, India, "Atmospheric Environment 39.40(2005):7745-India." Atmospheric Environment 39.40 (2005): 7745-7760.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details