IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.918 ⁜ Vol. 12, Issue 2, February 2023

DOI: 10.17148/IJARCCE.2023.12240

INDIAN AIR QUALITY PREDICTION ANDANALYSES USING ML

RAJASHEKHAR S A[1], SPOORTHI M[2], NISCHITHA P B[3], MANJUSHREE B H[4], SUSHMA M G[5]

Assistant Professor, CSE East West Institute of Technology, Bangalore, India[1]

Student, CSE, East West Institute of Technology, Bangalore, India[2-5]

Abstract: Air pollution includes things like dangerous gases and tiny particulate matter (PM2.5) that deteriorate air quality. This has developed into a crucial area for scientific research and a significant social issue that has an impact on the lives of the general public. As a result, multiple experts and academics at various R&D centres, institutions, and elsewhere are conducting extensive research on PM2.5 pollutant predictions. The authors provided a range of machine learning methods, such as linear regression and random forest models, in this scenario to forecast PM2.5 pollutants in polluted cities. This experiment is carried out with Python 3.7.3 and Jupyter Notebook. Observed to be more dependable models are random forest and based on results for the MAE, MAPE, and RMSE metrics among the models.

Keywords: Pollution detection, Pollution prediction, Logistic regression, Linear regression, Auto regression

I. INTRODUCTION

Due to rising air pollution, which is now a major issue in many regions of the world, accurate air pollution prediction and forecasting have become difficult and important tasks. Typically, there are two categories of pollution:

(1) natural pollution because of volcanic eruptions and forest fires resulting in emission of SO2, CO2, CO, NO2, and sulfate as air pollutants and

(2) man-made pollution because of some human activities such as burning of oils, discharges from industrial production processes, and transportation emissions that have PM2.5.

Due to its principal air pollutant, PM2.5, which has negative effects on the environment, other types of life, and human health, it has garnered a lot of attention. Many studies show that air pollution causes respiratory and cardiovascular diseases, which in turn cause the deaths of animals and plants, acid rain, climate change, global warming, etc., making it impossible for societies to function economically and for people to exist. According to Ameer et alanalysis .'s of the effects of PM2.5 over the past 25 Hindawi Adsorption Science & Technology using comparative analysis of ML techniques, 4.2 million people are thought to have died from prolonged exposure to PM2.5 in the atmosphere, with an additional 250,000 deaths attributed to ozone exposure. In terms of mortality risk factors, PM2.5 came in fifth place globally and was responsible for 7.6% of all fatalities. More than 20% of the 1.1 million deaths worldwide from 1990 to 2015 were linked to respiratory disorders, with China and India seeing the greatest increases. In order to better effectively regulate air pollution, a vast amount of research has been conducted globally on subjects including air pollution levels and air quality forecasts. Extensive research specifies that air pollution forecasting approaches can be imprecisely divided into three traditional classes:

(1) statistical forecasting methods,
(2) artificial intelligence methods
(3) numerical forecasting methods

When discharged into the atmosphere, PM2.5 pollutants—fine particles made up of a combination of harmful gases and particles—can cause harm

These pollutants are mainly responsible for causing human respiratory diseases in one way or another, and when if it is severe, it may also trigger the COVID-19 pandemic, which would increase the fatality rate. The survey clearly shows that PM2.5 causes more problems for people than other pollutants and that it is the one that generates other pollutants, hence the current models solely consider PM2.5 pollution. Using historical meteorological datasets, statistical analysis

for PM2.5 pollutant prediction is conducted. Few models are utilised for predicting, however the results indicated low error rate performance; the limitations of the available models prevent them from using several fundamental standard categorization approaches. In this proposed approach random forest model (RF) haS been implemented to predict the PM2.5 pollutant using meteorological and PM2.5 pollutant historical datasets that are downloaded from 1st Jan 2014 to 1st Dec 2019. These data have been monitored continuously for 24 h with a time period of an hour using the following

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified ⁜ Impact Factor 7.918 ⁜ Vol. 12, Issue 2, February 2023
DOI: 10.17148/IJARCCE.2023.12240

meteorological features such as temperature (T in ° C), minimum temperature (Tm in ° C), maximum temperature (TM in ° C), total train/snowmelt (PP in mm), humidity (H in %), wind speed (V in km/h), visibility (VV in km), and maximum sustained wind speed (VM in km/h). Also, the proposed machine learning models have been evaluated using statistical metrics such as Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), Mean Square Error (MSE), Root Mean Square Error (RMSE), and R2. Results show the achievement of better performance with decreased error rate when compared to traditional prediction models.

II. RELATED WORK

In this case, authors used the network's average residual battery level, which was computed by adding two fields to the RREQ packet header of an on-demand routing method.

i) average residual battery energy of the nodes on the path
ii) number of hops that the RREQ packet has passed through.

Retransmission time is proportional to remaining battery energy, according to their equation. Because their retransmission time is shorter, nodes with more battery energy than the average energy will be chosen. When the majority of nodes have the same retransmission time, a small hop count is chosen. A node's individual battery power is taken into account as a statistic to increase the network longevity. The authors employed an optimization function that takes into account the kind of packet, its size, the distance between nodes, the number of hops, and transmission time. The multicast group, which contains a number of paths from source to destination and the estimated lifetime of each path, was used to construct the initial population for the genetic algorithm. The path's lifetime is utilised as a fitness metric. The highest chromosomes with the longest lifetimes will be chosen using the fitness function. To improve the selection, cross over and mutation operators are utilized. By incorporating the concept of balanced energy usage into the route discovery process, authors improved the AODV protocol. When the nodes have enough energy to transport the message, the RREQ message will be transmitted; otherwise, the message will be dropped. This condition will be examined using a threshold value that changes dynamically. In order to extend the lifespan of the network, it enables a node with an overextended battery to refuse to route traffic. Authors added a power factor field to the AODV route table. The remaining nodes can be idle while only active nodes can choose the routing. Together with Hello packets, the lifetime of a node is estimated and sent. The authors took into account the node's individual battery capacity as well as the hop count, as a high hop count will aid to reduce the transmission power's range. Route discovery has been carried out similarly to how on-demand routing algorithms are carried out. Once a packet has arrived at its destination, it will wait for time t before gathering all of the packets. It calls the optimization algorithm to choose the path and send RREP after time t. The energy of each node is used by the optimization function; a node with low energy level will not be used by the optimization function.

III. PROPOSED ALGORITHM

A. Design Considerations:

- Random forest algorithm
- Mq3 sensor : is used to detect the presence of alcohol gases
- Node Mcq : is across platform
- LEDs : An electrical current passes through microchip which illuminates the tiny light sources we call LEDs
- Jupyter notebook in : is used it is an original web application
- Python 3.7.3

B. Description of the Proposed Algorithm:

Popular machine learning algorithm Random Forest is a part of the supervised learning methodology. It can be applied to ML issues involving both classification and regression. It is built on the idea of ensemble learning, whichis a method of integrating various classifiers to address difficult issues and enhance model performance.

Random Forest, as the name implies, is a classifier that uses a number of decision trees on different subsets of the provided dataset and averages them to increase the dataset's predictive accuracy. Instead than depending on a single decision tree, the random forest uses forecasts from each tree and predicts the result based on the votes of the majority of predictions.

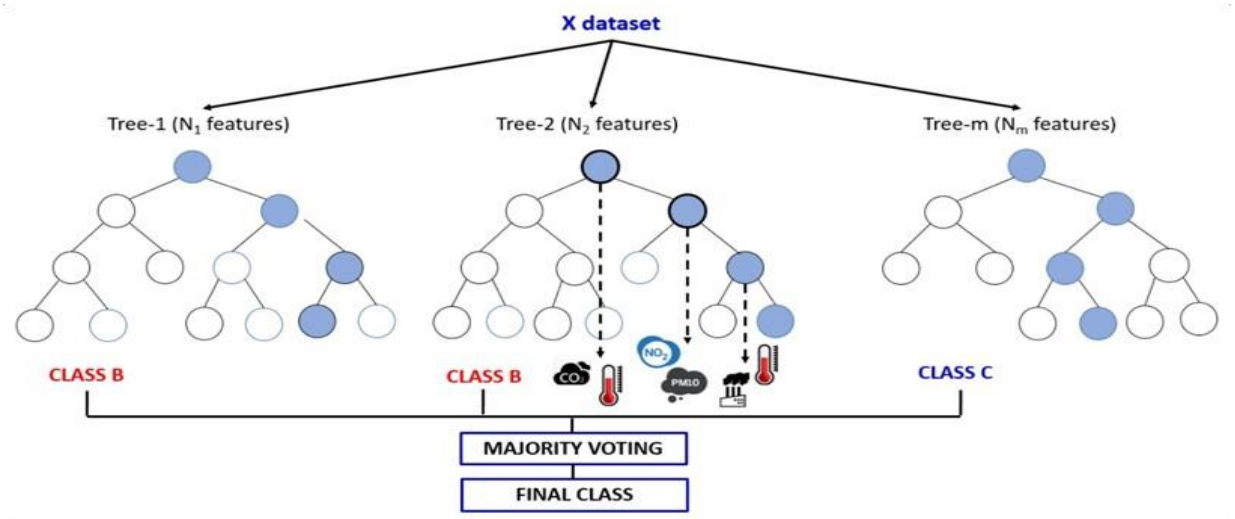
More trees in the forest result in increased accuracy and mitigate the overfitting issue.Some decision trees may predict the correct output, while others may not, because the random forest combines numerous trees to forecast the class of the dataset. But when all the trees are combined, they forecast the right result. As a result, the following twohypotheses for an improved Random forest classifier

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.918 ⁜ Vol. 12, Issue 2, February 2023

DOI: 10.17148/IJARCCE.2023.12240



In order for the classifier to predict accurate results rather than an assumed outcome, there should be some real valuesin the feature variable of the dataset. Each tree's predictions must have extremely low correlations.

IV. SIMULATION RESULTS

$$AQI = \begin{cases} max\{I_{O3}, I_{PM2.5}, I_{PM10}, I_{CO}, I_{SO2}, I_{NO2}\}, I_{O3}, I_{PM2.5}, I_{PM10} \neq \emptyset \\ \emptyset, otherwise \end{cases} \quad (1)$$

Pollutant concentration ($value_i$) is converted to pollutant index (I_i) by the following formula:

$$I_i = LB_i + \frac{value_i - lb_i}{ub_i - lb_i} \times UB_i - LB_i \quad (2)$$

Fig.1. Ad Hoc Network of 5 Nodes Fig. 2. Energy Consumption by Each Node Fig. 3. Ad Hoc Network of 5 Nodes Fig 4. Energy Consumption by Each Node

The hourly AQI data for the six pollutants O3, PM2.5, PM10, CO, SO2, and NO2 with the highest index are chosen as characteristics for the predictive models. The AQI Taiwan Guidelines [18] are used to convert the time-window-specific concentration of six pollutants, and the AQI is manually calculated using Equations (1) and (2). Index values of O3, PM2.5, and PM10 are required to define AQI in Taiwan, and the absence of one or more of these values will significantly decrease the accuracy of the assessment of the current air quality.

where i = O3, PM2.5, PM10, CO, SO2, NO2; j denotes which level in AQI system occupied by the concentration of the specific pollutant using categories of good, moderate, unhealthy which includes specific groups, unhealthy, very unhealthy, and hazardous. To calculate Ii values, the data transformation specifies the time-window-specific concentration. For instance, using the AQI from the Taiwan EPA website, the concentration valueO3 = 0.06 ppm will fall in the interval with lbO3 = 0.055 ppm and ubO3 = 0.070 ppm corresponding to the “moderate” pollutant level with LB moderate = 51 and UB moderate= 100. The value O3 is determined by either of two conditions being met: if the 8-h average concentration is lower than 0.2 ppm for a particular site and is also more precautionary, then this value is used; otherwise, the 1-h average concentration will be taken into account. Values PM2.5 and PM10 are moving averages that take into account the previous 12 and recent 4 hours, respectively (see Table 1). Some variables, like value CO and value NO2, only take into consideration the most recent 8 and 1 hours, respectively, of the time window. ValueSO2 emphasises the 24-h average concentration instead of the 1-h average concentration unless the 1-h average

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified ⁜ Impact Factor 7.918 ⁜ Vol. 12, Issue 2, February 2023
DOI: 10.17148/IJARCCE.2023.12240

concentration is greater than 185 ppb.

V. CONCLUSION AND FUTURE WORK

The effects of air pollution are detrimental to both the environment and human life. Air pollution occurs when the concentration of particular compounds in the atmosphere exceeds a certain threshold. Predicting PM2.5 and air quality is one of the most effective ways to reduce pollution. In the proposed models, the PM2.5 pollutant is predicted using meteorological datasets and six different models (LR, RF, KNN, RL, Xgb, and Adab models) are used for forecasting air quality levels. The results were evaluated using statistical metrics such as MAE, MAPE, MSE, RMSE, and R2. The better performance results for correlation coefficient determination in terms of R2 are KNN train set and test set values of 1.0 and -0.228, respectively; Xgb train set and test set values of 0.999 and 0.3072, respectively; and RF train set and test set values of 0.904 and 0.382, respectively. Among those proposed models from the results with respect to MAE, MAPE, and RMSE it could be obvious that Xgb, Adab, KNN, and RF are reliable models when compared to the existing models.

REFERENCES

[1] M. Castelli, F. M. Clemente, A. Popoviˇc, S. Silva, and L. Vanneschi, “A machine learning approach to predict air quality in california,” Complexity, vol. 2022
[2] V. Kanawade, A. Srivastava, K. Ram, E. Asmi, V. Vakkari, V. Soni, V. Varaprasad, and C. Sarangi, “What caused severe air pollution episode of november 2016 in new delhi?” Atmospheric Environment, vol.222, p. 117125, 2020
[3] D. C. Payne-Sturges, M. A. Marty, F. Perera, M. D. Miller, M. Swanson, K. Ellickson, D. A. Cory-Slechta, B. Ritz, J. Balmes, L. Anderko etal., “Healthy air, healthy brains: advancing air pollution policy to protect children’s health,” American journal of public health, vol. 109, no. 4, pp. 550–554, 2020.
[4] Y. Xu and H. Liu, “Spatial ensemble prediction of hourly PM2.5 concentrations around Beijing railway station in China,” Air Quality, Atmosphere and Health, vol. 13, no. 5, pp. 563–573, 2020.
[5] D. Kim, S. Cho, L. Tamil, D. J. Song, and A. S. Seo, “Predictingasthma attacks: effects of indoor PM concentrations on peak expiratory flow rates of asthmatic children,” IEEE Access, vol. 8, pp. 8791–8797, 2020.
[6] Gokhale sharad and Namita Raokhande, “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.
[7] Sivacoumar R, et al, “ Air pollution modelling for an industrial complex and model performance evaluation “, Environmental Pollution 111.3 (2010) : 471-477
[8] Singh Kunwar P., Shikha Gupta and Premanjali Rai, “ Identifying pollution sources and prediction urban air quality using ensemble learning methods”, Atmospheric environment80 (2013): 426-437.
[9] Carbajal-Hernández, José Juan "Assessment and prediction of air quality using fuzzy logic and autoregressive models." Atmospheric Environment 60 (2012): 37-50.
[10] Dragomir, Elia Georgiana. "Air quality index prediction using K-nearest neighbor technique no. 1 (2010): 103-108