INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

e-ISSN : 2583-1062

Impact Factor : 5.725

www.ijprems.com
editor@ijprems.com

Vol. 04, Issue 01, January 2024, pp : 490-496

IMPACT OF MACHINE LEARNING ON AIR QUALITY PREDICTION

Akshit Arora[*1]

[*1]Institute of Innovation in Technology and Management, India.

DOI: https://www.doi.org/10.58257/IJPREMS32575

ABSTRACT

Machine learning (ML) has emerged as a powerful tool in environmental science, particularly in predicting and monitoring air quality. This abstract explores the impact of machine learning on air quality prediction, highlighting its contributions, challenges, and implications. ML algorithms, including regression models, neural networks, and ensemble methods, have demonstrated significant improvements in accuracy and efficiency compared to traditional modeling approaches. The integration of diverse data sources, such as satellite imagery, meteorological data, and real-time sensor readings, allows ML models to capture complex relationships and dynamic patterns in air pollutant concentrations. Moreover, the ability to adapt and learn from new data enables continuous refinement of predictions, enhancing the reliability of air quality forecasts. Challenges include the need for high-quality labeled datasets, interpretability of complex models, and the ethical considerations associated with biased predictions. Despite these challenges, the application of ML in air quality prediction holds immense promise for advancing our understanding of pollution dynamics, informing timely interventions, and ultimately contributing to the mitigation of air quality-related health risks. As research in this field progresses, the collaboration between data scientists, environmental experts, and policymakers becomes crucial to harness the full potential of machine learning for improving air quality management and public health outcomes.

Keywords: Machine learning, air quality, quality management, health, data.

1. INTRODUCTION

In recent years, the field of air quality prediction has witnessed a transformative impact through the integration of machine learning (ML) techniques. As air pollution continues to be a pressing global concern with far-reaching implications for public health and the environment, the application of ML in predicting air quality has emerged as a promising avenue for enhancing the accuracy and efficiency of forecasting models.

Traditional air quality prediction models often rely on complex mathematical formulations and simulations that may struggle to capture the intricate dynamics of pollutant concentrations. Machine learning, with its capacity to discern patterns and relationships from vast datasets, offers a paradigm shift in how we approach air quality prediction. This introduction delves into the significant impact of machine learning on air quality prediction, shedding light on the key advancements, challenges, and potential implications for environmental management.

Advancements in Accuracy and Precision

Machine learning algorithms, ranging from regression models to advanced neural networks and ensemble methods, have demonstrated remarkable capabilities in improving the accuracy and precision of air quality predictions. By harnessing historical data on pollutant concentrations, meteorological conditions, and other relevant variables, these models can discern intricate patterns that may be challenging for traditional models to capture. The result is more nuanced and reliable predictions, allowing for a deeper understanding of pollution dynamics.

Integration of Diverse Data Sources

One of the defining features of ML in air quality prediction is its ability to integrate diverse data sources. Beyond relying solely on ground-based monitoring stations, ML models incorporate information from satellite imagery, meteorological data, and real-time sensor readings. This holistic approach enables a comprehensive analysis of contributing factors, providing a more complete and real-time assessment of air quality conditions.

Continuous Learning and Adaptability

Unlike static models, ML algorithms exhibit the capacity for continuous learning and adaptability. As new data becomes available, these models can dynamically adjust and refine their predictions, ensuring that the forecasts remain accurate and up-to-date. This adaptability is particularly crucial in the context of rapidly changing environmental conditions and evolving pollution sources.

Challenges and Considerations

While the impact of ML on air quality prediction is undeniable, challenges persist. The need for high-quality labeled datasets, interpretability of complex models, and ethical considerations related to potential biases in predictions are

among the key challenges that researchers and practitioners must address. Overcoming these hurdles is essential to fully harness the potential of ML in the context of air quality management.

Implications for Environmental Management

The application of ML in air quality prediction holds significant implications for environmental management and public health. Accurate and timely predictions enable policymakers to implement proactive measures, inform the public about potential risks, and allocate resources efficiently. The collaboration between data scientists, environmental experts, and policymakers becomes paramount in maximizing the benefits of ML for improving air quality and mitigating the adverse effects of pollution.

As we delve into the intricate interplay between machine learning and air quality prediction, this exploration seeks to unravel the transformative potential of ML in revolutionizing our understanding of air pollution dynamics and, ultimately, contributing to a healthier and more sustainable environment.

2. REVIEW OF LITERATURE

Castelli et al. (Using the Support Vector Regression (SVR) machine learning algorithm, 2020) attempted to forecast California's air quality in terms of pollutants and particulate levels. The creators professed to foster a clever technique to display hourly air contamination. Doreswamy and co. 2020) researched ML prescient models for determining PM fixation in the air. The creators concentrated on six years of air quality observing information in Taiwan and applied existing models. They asserted that actual and predicted values were very close to one another. Liang and others 2020) concentrated on the exhibitions of six ML classifiers to foresee the AQI of Taiwan in light of 11 years of information. The creators announced that Versatile Helping (AdaBoost) and Stacking Gathering are generally reasonable for air quality expectation yet the anticipating execution fluctuates over various geological locales. Madan et al. (2020) analyzed twenty different artistic works over poisons examined, ML calculations applied, and their individual exhibitions. The creators found that many works integrated meteorological information, for example, moistness, wind speed, and temperature to anticipate contamination levels all the more precisely. They discovered that NN and boosting models performed better than the other well-known ML algorithms. Madhuri et al. (Wind speed, wind direction, humidity, and temperature all had a significant impact on the concentration of pollutants in the air, according to 2020). The creators utilized managed ML procedures to foresee the AQI and found that the RF calculation showed the least order blunders. Monisri et al. (2020) gathered air contamination information from different sources and tried to foster a blended model for foreseeing air quality. The creators guaranteed that the proposed model plans to assist with peopling in humble communities to dissect and anticipate air quality. Nahar et al. (2020) fostered a model to foresee AQI in light of ML classifiers. The authors looked at the data that was collected by Jordan's ministry of environment over a 28-month period and found the concentrations of pollutants. Their proposed model recognized the most sullied regions with fulfilling exactness. Patil et al. (2020) introduced a few scholarly deals with different ML methods for AQI demonstrating and estimating. The majority of researchers used Artificial Neural Network (ANN), Linear Regression (LR), and Logistic Regression (LogR) models to predict AQI, according to the authors.

Bhalgat et al. (2019) applied the ML method to foresee the convergence of SO2 in the climate of Maharashtra, India. The creators reasoned that being profoundly dirtied, a few urban communities of this Indian territory require grave consideration. The creators referenced that their model was not fit for displaying anticipated yields. Mahalingam et al. (2019) fostered a model to foresee the AQI of brilliant urban communities and tried it in Delhi, India. The creators detailed that the medium Gaussian Help Vector Machine (SVM) displayed greatest exactness. The creators guarantee that their model can be utilized in other shrewd urban areas as well. Soundari et al. (2019) fostered a model in light of NNs to foresee the AQI of India. The creators asserted that their proposed model could foresee the AQI of the entire district, of any territory, or of any geological locale when the previous information on grouping of poisons were accessible.

Sweileh et al. (2018) concocted an exceptionally intriguing learn about the examination of worldwide companion evaluated writing about air contamination and respiratory wellbeing. The creators separated 3635 reports from the Scopus data set distributed somewhere in the range of 1990 and 2017. They saw that there was a significant expansion in distributions from 2007 to 2017. The creators detailed dynamic nations, organizations, diaries, creators, global joint efforts in the domain and reasoned that exploration deals with air contamination and respiratory wellbeing had been getting a ton of consideration. They recommended getting popular feelings about moderation of open air contamination and interest in green advances. Zhu et al. (2018) refined the issue of AQI forecast as a perform various tasks learning issue. The creators used enormous scope advancement procedures and attempted to decrease the

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

e-ISSN : 2583-1062

www.ijprems.com
editor@ijprems.com

Vol. 04, Issue 01, January 2024, pp : 490-496

Impact Factor : 5.725

quantity of boundaries. In view of their experimental outcomes, they guaranteed that the proposed model showed improved results than existing relapse models.

Bellinger and others 2017) did an itemized writing investigation on the utilization of ML and information mining techniques toward air contamination the study of disease transmission. The creators tracked down that the scientists from Europe, China, and the USA were extremely dynamic in this domain and the accompanying classifiers had been generally applied: Choice Tree (DT), SVMs, K-implies grouping, and the APRIORI calculation. Rybarczyk and Zalakeviciute (2017) tried to foster a model that related traffic thickness with air contamination. The creator referenced that such traffic information assortment was practical, and coordinating it with meteorological highlights helped precision. The creators viewed that as the half and half model played out the best and exactness in light of morning time information was the most noteworthy.

3. METHODOLOGY

Some Indian cities fall in the array of the most polluted cities in the world, and the threat of air pollution is being raised day by day. Poor air quality in India is now considered a significant health challenge and a major obstacle to economic growth. According to a new study released jointly by a UK-based non-profit management firm, *Dalberg Advisors and Industrial Development Corporation*, air pollution in India caused annual losses of up to Rs 7 lakh crore ($95 billion) (Dalberg 2019). The main pollutant emissions in India are due to the energy production industry, vehicle traffic on roads, soil and road dust, waste incineration, power plants, open waste burning, etc. The present research investigates air pollution data extracted from the *Central Pollution Control Board (CPCB)*, India.

Data preprocessing

Nature of information is the first and generally significant essential for powerful representation and production of proficient ML models. The preprocessing steps assist in diminishing the commotion with introducing in the information which at last speeds up and speculation ability of ML calculations. Anomalies and missing information are the two most normal mistakes in information extraction and observing applications. The information preprocessing step performs different procedure on information, for example, finishing up not-a-number (NAN) information, eliminating or changing exception information, and so on. Figure 2 displayed underneath presents a perspective on the missing qualities in each element of the dataset. See that among any remaining highlights, Xylene has the most missing qualities and CO has the most un-missing qualities. Countless missing qualities might be existing because of various elements, for example, a station that can detect information yet doesn't have a gadget to record it.

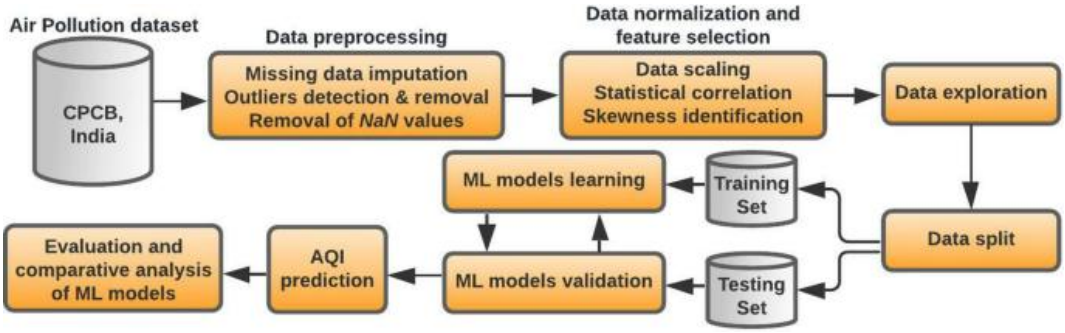

Fig. 1 : Flowchart of the proposed model

	Missing Values	% of Total Values
Xylene	18109	61.300000
PM10	11140	37.700000
NH3	10328	35.000000
Toluene	8041	27.200000
Benzene	5623	19.000000
AQI	4681	15.900000
AQI_Bucket	4681	15.900000
PM2.5	4598	15.600000
NOx	4185	14.200000
O3	4022	13.600000
SO2	3854	13.100000
NO2	3585	12.100000
NO	3582	12.100000
CO	2059	7.000000

Fig. 2 : Missing values of the features and their percentages

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

www.ijprems.com
editor@ijprems.com

Vol. 04, Issue 01, January 2024, pp : 490-496

e-ISSN : 2583-1062

Impact Factor : 5.725

Every one of the missing qualities are loaded up with the middle qualities against each element to take care of the missing information issue. Then, a standardization cycle has been applied to normalize the information, guaranteeing that the meaning of factors is unaffected by their reaches or units. The information standardization process assists with bringing various information credits into a comparative size of estimation. This cycle assumes a fundamental part in the steady preparation of ML models and lifts execution.

4. ANALYSIS

When data have a normal distribution, many machine learning models perform better, but when data have a skewed distribution, they perform poorly. Hence, it is important to distinguish the skewness being available in the highlights and to play out certain changes and mappings which convert the slanted circulation into an ordinary conveyance. The characteristics of CO, Xylene, Benzene, and Toluene are highly skewed, as shown in Figure 4. To make these slanted elements more typical, the logarithmic changes have been utilized to decrease the effect of anomalies by normalizing size contrasts.

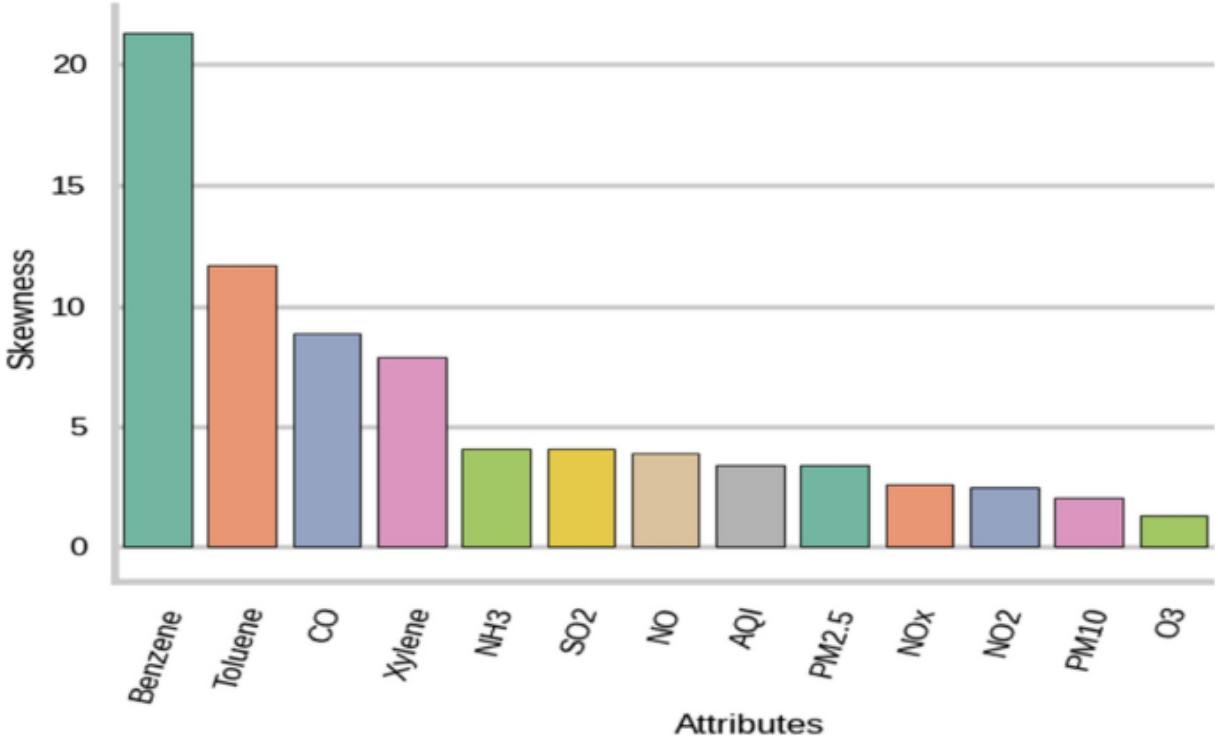

Fig. 3 : Skewness present in dataset features

Analyzing the patterns of air pollutants, India has emerged as one of the few nations with the most severe air pollution over the past few years due to rapid industrialization and rapid urbanization. Air contamination is among grave general wellbeing and ecological issues, and the Wellbeing Impacts Organization (HEI) positions it among the best five worldwide gamble factors for mortality (IHME 2019). The HEI study found that PM emissions were the third leading cause of death in 2017, and India had the highest rate. The World Health Organization (WHO) ranked India as the fifth most polluted nation based on the emissions of PM2.5 and other pollutants (Gurjar, 2021). From 2015 to 2020, various pollutants' trends are observed and depicted in the figure below (Fig. 4). See that with the exception of O3 and Benzene, any remaining poisons showed a critical fall in 2020. The year 2020 saw the most severe lockdown throughout the entire existence of humankind and stopped modern, vehicle, and flight exercises in India and the world filled in as a few ambrosia for the sickly climate and air.

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

e-ISSN : 2583-1062

Impact Factor : 5.725

www.ijprems.com
editor@ijprems.com

Vol. 04, Issue 01, January 2024, pp : 490-496

Fig. 4 : Intensities of various pollutants from 2015 to 2020

Toxins that are straightforwardly engaged with expanding AQI values

The relationship values between various toxins and AQI have been practiced and the poisons for which this connection esteem is more noteworthy than the limit of 0.5, for example the relationship is unequivocally certain have been recognized. Figure 5 displayed underneath portrays the convergence of four such toxins in different urban areas in India.

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

e-ISSN : 2583-1062

Impact Factor : 5.725

www.ijprems.com
editor@ijprems.com

Vol. 04, Issue 01, January 2024, pp : 490-496

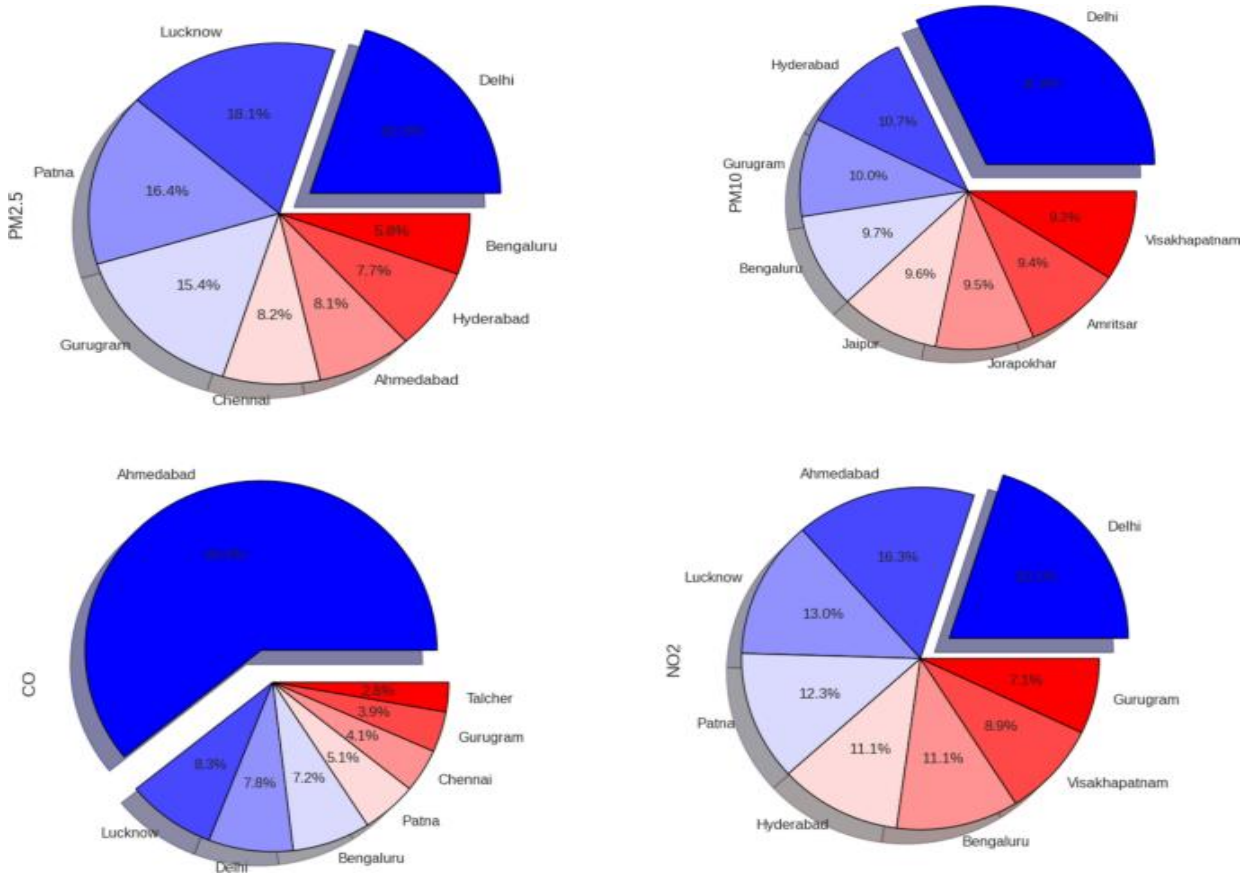

Fig. 5 : Pollutants governing AQI directly

5. CONCLUSION

Forecast of air quality is a difficult undertaking in view of the unique climate, unconventionality, and changeability in existence of toxins. The grave outcomes of air contamination on people, creatures, plants, landmarks, environment, and climate call for steady air quality checking and examination, particularly in agricultural nations. Notwithstanding, lesser consideration for scientists has been noticed for AQI expectation for India. In the current work, air contamination information of 23 Indian urban communities for a residency of six years are examined. The dataset is cleaned and preprocessed first by filling NAN values, tending to anomalies, and normalizing information values. Then, at that point, relationship based highlight choice method is practiced to channel AQI influencing poisons for additional review and logarithmic changes are applied to the slanted elements. Utilizing exploratory data analysis techniques, various hidden patterns in the dataset are discovered. It was found that practically all contaminations showed a critical fall in 2020. The information irregularity issue is tended to by the Destroyed investigation. A ratio of 75–25% divides the dataset into train and test subsets. ML-based AQI forecast is done with and without Destroyed resampling method and a near examination is introduced. The consequences of ML models for both the train-test subsets are introduced as far as standard measurements like exactness, accuracy, review, and F1-Score. For both the train-test sets, the XGBoost model accomplished the most elevated exactness and the SVM model showed the least precision. The traditional factual mistake measurements, in particular MAE, RMSE, RMSLE, and R2 are then assessed to survey and think about the exhibitions of ML models. The XGBoost model emerges to be the general best entertainer by accomplishing the ideal qualities in both preparation and testing stages. When exercised with SMOTE, the RF model performed fairly well during the training phase. Then again, practically all ML models displayed upgrades in the testing stage. The GNB model achieved the best target prediction results for R2 during this phase. The current examination attempts to add to the writing by tending to air quality investigation and expectation for India which could have not been as expected considered. This work can be reached out by utilizing profound learning methods for AQI expectation.

6. REFERENCES

[1] Bellinger C, Jabbar MSM, Zaïane O, Osornio-Vargas A (2017) A systematic review of data mining and machine learning for air pollution epidemiology. BMC Public Health. https://doi.org/10.1186/s12889-017-4914-3

[2] Bhalgat P, Bhoite S, Pitare S (2019) Air Quality Prediction using Machine Learning Algorithms. Int J Comput Appl Technol Res 8(9):367–370. https://doi.org/10.7753/IJCATR0809.1006

[3] Castelli M, Clemente FM, Popovi˘c A, Silva S, Vanneschi L (2020) A machine learning approach to predict air quality in California. Complexity 2020(8049504):1–23. https://doi.org/10.1155/2020/8049504

[4] Doreswamy HKS, Yogesh KM, Gad I (2020) Forecasting Air pollution particulate matter (PM2.5) using machine learning regression models. Procedia Comput Sci 171:2057–2066. https://doi.org/10.1016/j.procs.2020.04.221

[5] Liang Y, Maimury Y, Chen AH, Josue RCJ (2020) Machine learning-based prediction of air quality. Appl Sci 10(9151):1–17. https://doi.org/10.3390/app10249151

[6] Madan T, Sagar S, Virmani D (2020) Air quality prediction using machine learning algorithms–a review. In: 2nd international conference on advances in computing, communication control and networking (ICACCCN) pp 140–145. https://doi.org/10.1109/ICACCCN51052.2020.9362912

[7] Madhuri VM, Samyama GGH, Kamalapurkar S (2020) Air pollution prediction using machine learning supervised learning approach. Int J Sci Technol Res 9(4):118–123

[8] Mahalingam U, Elangovan K, Dobhal H, Valliappa C, Shrestha S, Kedam G (2019) A machine learning model for air quality prediction for smart cities. In: 2019 international conference on wireless communications signal processing and networking (WiSPNET). IEEE 452–457. https://doi.org/10.1109/WiSPNET45539.2019.9032734

[9] Monisri PR, Vikas RK, Rohit NK, Varma MC, Chaithanya BN (2020) Prediction and analysis of air quality using machine learning. Int J Adv Sci Technol 29(5):6934–6943

[10] Nahar K, Ottom MA, Alshibli F, Shquier MA (2020) Air quality index using machine learning—a jordan case study. COMPUSOFT, Int J Adv Comput Technol 9(9):3831–3840

[11] Patil RM, Dinde HT, Powar SK (2020) A literature review on prediction of air quality index and forecasting ambient air pollutants using machine learning algorithms 5(8):1148–1152

[12] Rybarczyk Y, Zalakeviciute R (2017) Regression models to predict air pollution from affordable data collections. In: H. Farhadi (Ed.), Machine learning advanced techniques and emerging applications pp 15–48. IntechOpen. https://doi.org/10.5772/intechopen.71848

[13] Soundari AG, Jeslin JG, Akshaya AC (2019) Indian air quality prediction and analysis using machine learning. Int J Appl Eng Res 14(11):181–186

[14] Sweileh WM, Al-Jabi SW, Zyoud SH, Sawalha AF (2018) Outdoor air pollution and respiratory health: a bibliometric analysis of publications in peer-reviewed journals (1900–2017). Multidiscip Respiratory Med. https://doi.org/10.1186/s40248-018-0128-5

[15] Zhu D, Cai C, Yang T, Zhou X (2018) A machine learning approach for air quality prediction: model regularization and optimization. Big Data and Cognitive Comput. https://doi.org/10.3390/bdcc2010005