Forecasting India's Air Quality: A Machine Learning Approach for Comprehensive Analysis and Prediction

Nabin Kumar Sah, Osho Kothari, Konda V S Harshith Kumar

Department of Computer Science and Engineering

Amrita School of Computing, Bangalore, Amrita Vishwa Vidyapeetham, India

BL.EN.U4CSE21129@bl.students.amrita.edu, BL.EN.U4CSE21141@bl.students.amrita.edu,

BL.EN.U4CSE21101@bl.students.amrita.edu

***Abstract* — With rising industrialization, India confronts increasing difficulties in maintaining air quality regulations. This research proposes a comprehensive analysis and prediction framework based on machine learning approaches for assessing and forecasting the Air Quality Index (AQI) in various locations of India. Our model takes into account a wide variety of pollutant concentrations that are monitored at regular intervals. Our approach, which uses historical data from numerous Indian cities, employs a variety of machine learning methods, including k-Neighbour regression, decision tree regressor, support vector regressor, and random forest regressor, in addition to linear regression. Our model achieves an outstanding 99% accuracy versus the given dataset, indicating strong performance and a substantial progress in air quality prediction. The environmental health risks associated with air pollution have been mitigated by the addition of important knowledge.**

Key words: Air-pollution, Air quality Index, pollution, Indian air quality.

I. INTRODUCTION

India's recent decades have seen a rapid industrialization that has increased air pollution, posing a threat to human health and the environment. Robust procedures are needed to perceive, monitor, and forecast the country's air quality given the increasing concentrations of pollutants and other dangerous airborne particles.

Reducing the harmful effects of pollution requires careful monitoring and management of air quality. We monitor multiple contaminants, each with its own scale and index, so that we may accurately assess the condition of the air. The Air Quality Index (AQI) can be used to identify and understand amounts of major pollutants. We have used known Indian air quality guidelines to rigorously quantify and categorize pollutant concentrations, which forms the foundation of our machine learning-based analysis and prediction model.

Computing unique pollutant indices for each data point is the initial stage in our study process. The AQI values from various regions of India are compiled to replace this. AQI is successfully predicted by applying machine learning approaches, such as k-Neighbour regression, decision tree regression, support vector regression, and random forest regression, which leverage the power of historical data. The 99% accuracy rate of the model when evaluated against the given dataset demonstrates its usefulness. This research will help environmentalists, politicians, and the general public by filling in the knowledge gap concerning India's air quality between what is known and what may be projected.

To sum up, our study constitutes a significant advancement in the comprehension and prediction of air quality in India. We contribute to the greater goal of environmental sustainability and public health by combining state-of-the-art machine learning algorithms with extensive pollution data to produce a reliable tool for air quality index prediction.

II. RELATED WORK

In [1] the authors employ gradient descent boosted multivariable regression to forecast the Air Quality Index (AQI) in India, achieving 96% accuracy. Historical data is used, and the model surpasses standard regression models. The Analytic Hierarchy Process Multiple Criteria Decision Making (AHP MCDM) technique is used for preference order. Keywords include AQI, dataset preprocessing, outliers, BVA, and prediction. In [2], the author Addresses global pollution impact on cities like Delhi, Beijing, and Tehran using various machine learning algorithms. Random Forest Regression (RFR) outperforms other methods in predicting AQI values. In [3], they Focuses on environmental impact of technological advancements on air quality in India. Supervised machine learning algorithms are used to model and predict future AQI values, emphasizing the health risks associated with increased pollutant levels. In [4] the authors review recent advancements in air quality monitoring and forecasting, emphasizing the use of machine learning techniques. Provides a comprehensive overview of methodologies, merits, demerits, and identifies challenges in addressing air pollution.[5]: Highlights the significance of predicting air quality for public health, utilizing various machine learning algorithms. Despite ongoing research, accuracy is not entirely achieved. The study contributes to refining air quality prediction models using datasets from sources like Kaggle. [6]: Addresses air pollution with a comparative analysis of machine learning algorithms, demonstrating their appropriateness for predicting AQI using various pollutants. Offers insights into combating challenges posed

by air pollution.[7]: Focuses on establishing an accurate air quality prediction model using data mining methods and neural networks. Utilizes a Self-Organizing Map (SOM) and NSGA-II optimized neural network for predicting pollution scenarios with over 90% accuracy.
[8]: Emphasizes air quality monitoring in industrial and urban areas using big data analytics and machine learning. Reviews research results on air quality evaluation, incorporating artificial intelligence, decision trees, deep learning, and addresses challenges in air quality prediction.[9]: Focuses on monitoring and predicting air quality in India using machine learning techniques. Utilizes six years' worth of data from 23 cities, addressing data imbalance, and identifies XGBoost as a standout performer for accurate air quality prediction.[10]: Concentrates on predicting Particulate Matter with a diameter of less than 2.5µm (PM2.5) using deep learning. The hybrid CNN-LSTM model outperforms traditional models, contributing insights into precise air quality predictions, especially for PM2.5 in urban areas.[11]: Emphasizes the severe impact of air pollution on human life in India. Utilizes various machine learning algorithms to enhance the analysis of the AQI, addressing the complexities of air pollution.: Underscores the importance of examining air quality in industrial and urban areas. Focuses on predicting air pollution, particularly PM2.5, using data mining and machine learning techniques, contributing valuable insights into mitigating the impact of pollution in urban environments.

III. METHODOLOGY

In this segment, we elucidate the methodology employed for predicting and analysing air quality in India through the utilization of machine learning. Machine learning algorithms have become instrumental tools in forecasting air quality. Researchers employ a diverse set of techniques to scrutinize historical data, meteorological variables, and concentrations of pollutants.

3.1 Data Collection and Pre-processing

We obtained the air quality dataset from Kaggle, encompassing various features. This dataset, utilized in our study, comprises a diverse array of sensor data gathered from different locations throughout India. It offers valuable insights into ambient air quality, with a specific focus on crucial air quality parameters such as Sulphur dioxide, Nitrogen dioxide, Respirable Suspended Particulate Matter, and Suspended Particulate Matter.

Within the dataset, there exists an attribute named Pm_2.5, where 97% of the data is null. Despite being a major pollutant, it was deemed necessary to drop this column. The dataset originally had a shape of 435,742 instances and seventeen attributes.

Additionally, we identified attributes like std_code, sampling date, agency, location monitoring station, and the date on which the data was collected, which played no significant role in air quality prediction. Consequently, we decided to drop these irrelevant columns. Following this step, we proceeded to fill null numerical values for different pollutant concentrations with the mean value of the respective pollutant, and categorical values with null entries were filled using the mode.

Given the substantial size of the dataset, we also eliminated all instances where the SPM values were null. After completing these preprocessing steps, the new shape of the data frame stands at 198,355 rows and 8 columns.

	so2	no2	rspm	spm	pm2_5
count	401096.000000	419509.000000	395520.000000	198355.000000	9314.000000
mean	10.829414	25.809623	108.832784	220.783480	40.791467
std	11.177187	18.503086	74.872430	151.395457	30.832525
min	0.000000	0.000000	0.000000	0.000000	3.000000
25%	5.000000	14.000000	56.000000	111.000000	24.000000
50%	8.000000	22.000000	90.000000	187.000000	32.000000
75%	13.700000	32.200000	142.000000	296.000000	46.000000
max	909.000000	876.000000	6307.033333	3380.000000	504.000000

Fig1: Dataset description

3.2 Visualization

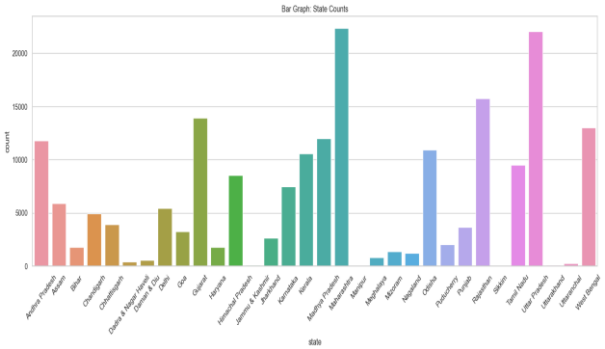

Fig2: Number of instances of different cities

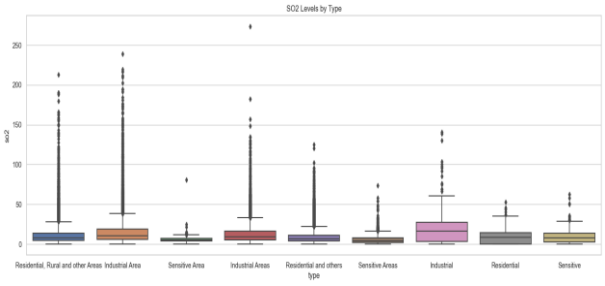

Fig3:So2 level in the different areas

Fig 1 is the dataset description of the numerical value of the different pollutants. Fig 2 is the bar graph that shows the number of instances that are present in different cities and the fig 3 is the so2 concentration present in different areas like residential areas, industrial areas and other areas.

3.3 AQI calculation: AQI is an acronym for Air Quality Index, representing a standardized measure of pollutant concentrations in the ambient air, providing a unified and easily interpretable AQI that reflects the potential health risks associated with varying pollution levels.
First, we have collected the major pollutant of the air then we have converted the concentration of each pollutant to the AQI sub- index using a specific formula. AQI value for India ranges between 0-500. The pollutant having highest

AQI value is considered the pollutant of the most concern at that time.

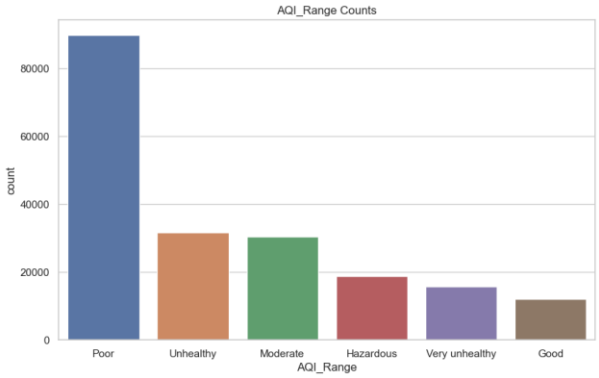

Fig 4: Count of instances having the AQI_Ranges.

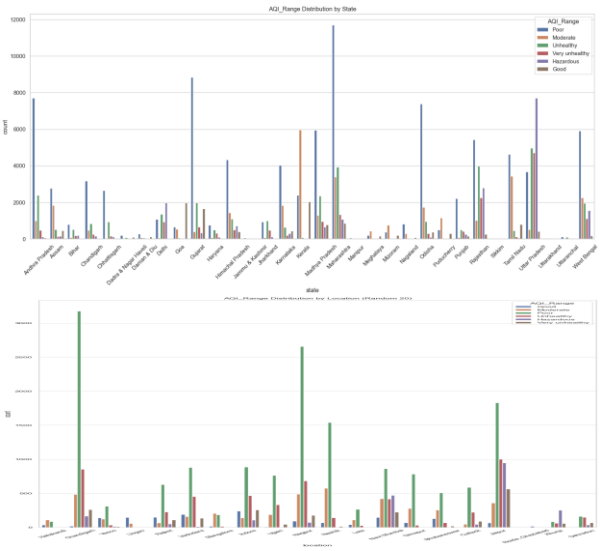

Fig 5: Different location cities of India showing the air quality index ranges Distribution.

Figure 5 is the visualization number of instances having the which level of the AQI value, In the given figure, we can see that the number of poor levels is highest in number which suggest that many of the cities are too much polluted and unhealthy. similarly graph 5 shows the number of instances with the AQI value Range in different states.

3.4 Feature scaling: first we have calculated the index having the outlier and the after removing the outlier we have applied the Normalization technique so that the dataset will be consistent and flexible.

3.5 Feature selection: From the pollutant provided in the dataset we have found the index for each pollutant and that pollutant are very necessary for the prediction. So, SOI, NOI, SPMI, and Rpi is taken for the features i.e., as X and for the target variable we have taken AQI, because we need to find the AQI if the index of the pollutant is given.

3.6 Training and Testing: In the process of training a machine learning model, it's common practice to use a substantial portion of the available data for training, and this is typically around 80%. This ensures that the model learns from a diverse set of examples, capturing patterns and relationships within the data. The approach commonly employed for this purpose is random sampling, which involves randomly selecting 80% of the data for the training phase. This method is widely accepted and proven to yield effective results in empirical studies**.**

4. Workflow of The Models

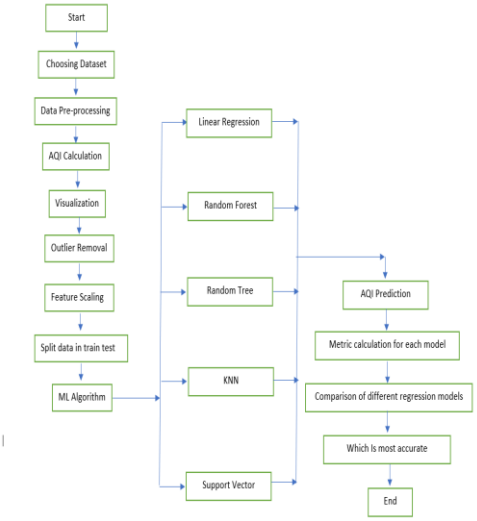

Fig: Flowchart of the proposed methodology

5. Machine Learning Models

5.1 Linear Regression: In machine learning, linear regression is a statistical technique employed to model the connection between one or more independent variables and dependent variable. The fundamental assumption is that this connection is linear, indicating that alterations in the independent variables exhibit a linear correlation with changes in the dependent variable. The key goal is to ascertain the optimal-fitting line that reduces the difference between the actual and the predicted value, enabling the prediction or understanding of the dependent variable based on the independent variable's values.

```
Linear Regression:
Mean Squared Error (MSE): 2.517859944030479e-05
Root Mean Squared Error (RMSE): 0.00501782815970264
R2 Score: 0.9994320970075244
Mean Absolute Error (MAE): 0.001141428761348404
```

Fig 6: Metrics of linear regression

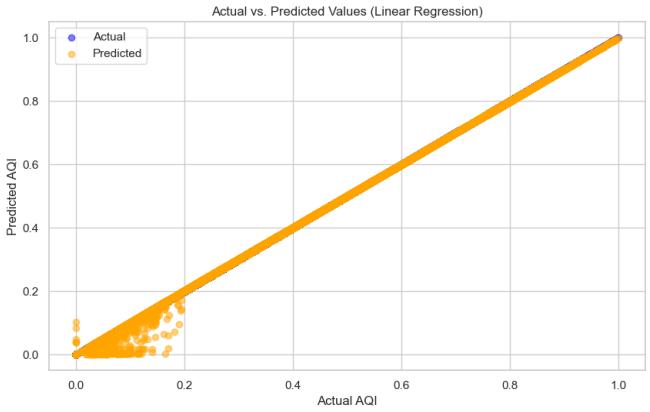

Fig 7: Actual and predicted value of the linear regression

5.2 KNN Regression:

The K-Nearest Neighbors (KNN) regression predicts continuous outcomes by taking the average of the target values from its nearest k neighbors into account. It is non-parametric, does not assume data distribution, and is easy to grasp. However, with large datasets, it can be noisy and computationally expensive. The parameter k is critical in influencing model sensitivity.

```
KNN Regression Metrics:
Mean Squared Error (MSE): 1.306459359034015e-05
Root Mean Squared Error (RMSE): 0.0036144976954398727
R2 Score: 0.9997053282565211
Mean Absolute Error (MAE): 0.0021793281423242725
```

Fig 8: Metrics of KNN regression

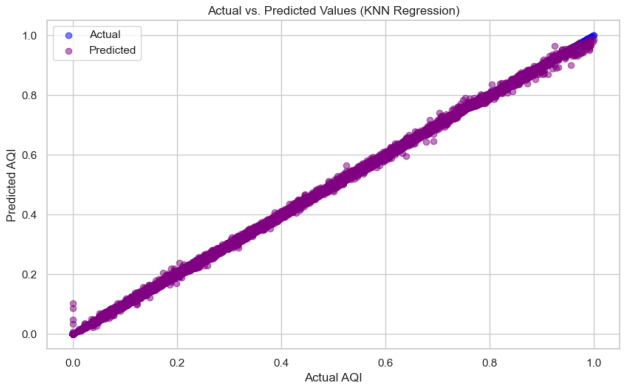

Fig 9: Actual and predicted value of KNN

5.3 Decision Tree Regression: Decision trees are well-known for their intuitive character, ease of comprehension, and ability to capture complicated data linkages. They provide a clear depiction of the decision-making process, making them especially useful for interpretation. However, one significant difficulty with decision trees is their proclivity to overfitting, which occurs when the model becomes overly tailored to the training data and struggles to generalize effectively to novel data. Pruning techniques, which include eliminating sections of the tree that do not significantly contribute to predictive accuracy, are commonly employed to resolve overfitting and enhance the model's performance on new data.

```
Decision Tree Regression Metrics:
Mean Squared Error (MSE): 2.576347314166328e-06
Root Mean Squared Error (RMSE): 0.0016051004062569818
R2 Score: 0.9999418905188575
Mean Absolute Error (MAE): 3.348324527586457e-05
```

Fig 10: Metrics of Decision Tree regression

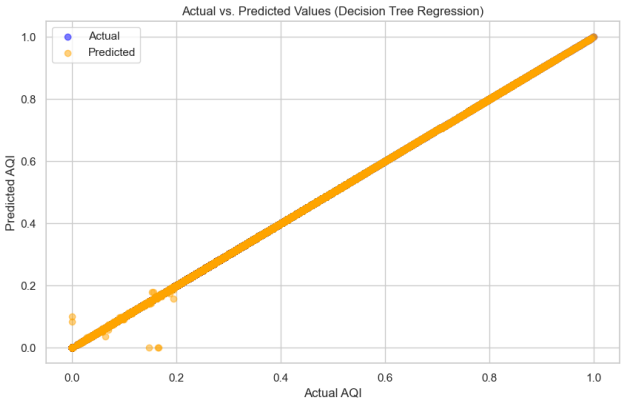

Fig 11: Actual and predicted value of Decision Tree Regression

5.4 Support vector Regression

SVR is a regression algorithm that uses support vector machines to accomplish regression tasks. It operates by selecting a hyperplane that best reflects the relationship between input and output variables, as well as incorporating an epsilon-tube tolerance margin. The major goal is to limit errors within this margin while still providing a good fit to the data. SVR is particularly adept at dealing with non-linear relationships, making it particularly useful in situations where traditional linear regression approaches may be insufficient for complex datasets.

```
SVR Metrics:
Mean Squared Error (MSE): 0.0006624668587116632
Root Mean Squared Error (RMSE): 0.02573843155111949
R2 Score: 0.9850580683659484
Mean Absolute Error (MAE): 0.019971514513380927
```

Fig 12: Metrics of Support vector regression

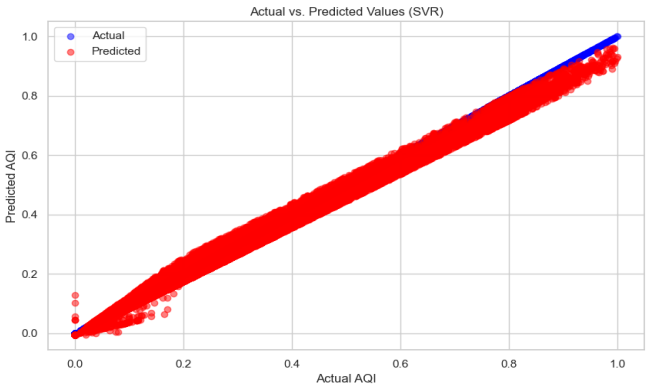

Fig 13: Actual and predicted value of Support vector Regression

5.5 Random Forest Regression:

Random Forest Regression is a machine learning technique created specifically for regression. During the training step, it generates several decision trees, resulting in a final prediction derived from the average output of these individual trees. This ensemble technique improves prediction accuracy and robustness, especially when dealing with complex connections between variables. Random Forest Regression, known for its versatility, ability to manage large datasets, and resistance to overfitting, has emerged as a popular approach for addressing regression difficulties in a variety of disciplines.

```
Random Forest Regression Metrics:
Mean Squared Error (MSE): 9.144257947173417e-07
Root Mean Squared Error (RMSE): 0.0009562561344730509
R2 Score: 0.9999793751377455
Mean Absolute Error (MAE): 2.830990984160992e-05
```

Fig 14: Metrics of Random Forest Regression

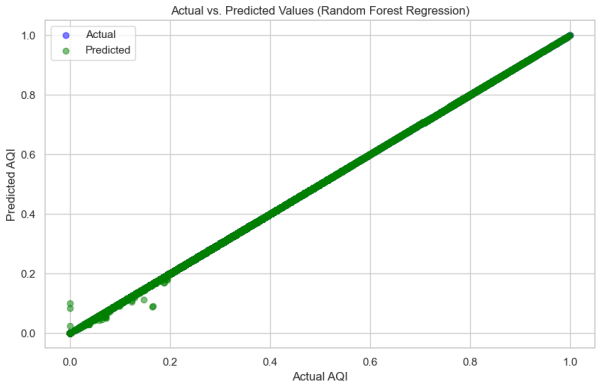

Fig 15: Actual and predicted value of Random Forest Regression

The scatter plots presented above indicates a favorable correlation between the anticipated and observed values, with a clustering of data points along the diagonal line. This suggests the effectiveness of the models in accurately predicting actual values. Notably, several algorithms exhibit R-squared values surpassing 0.98, indicating that the models capture approximately 98% of the variability in the dataset. As a result, it can be concluded that each machine learning regression model appears to be well-suited for the given dataset.

IV. RESULTS

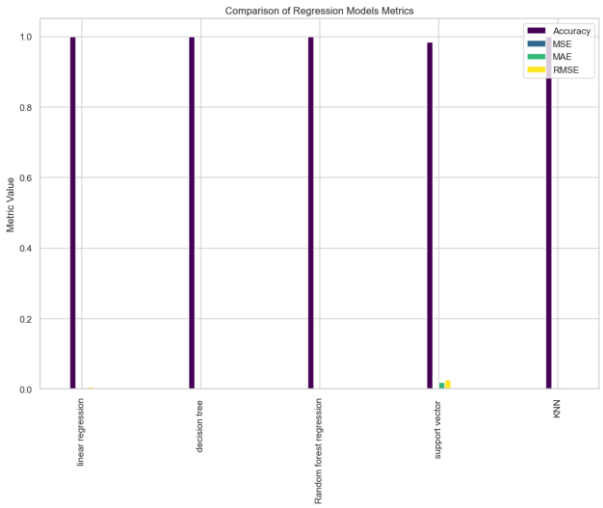

Fig 16: Metrics of different model

From the above table of the different regression model, we can see that the all the model that we have taken has the accuracy of 99% and 98%, while the other metrics have the low value. The other metrics having low values indicates the error of the actual and predicted value is low which is good for any of the Machine learning Model.

Comparison Metrics of Different Models

Models	R-squared	RMSE	MSE	MAE
Linear Regression	0.9994	0.00501	2.517e-05	0.00114
Decision tree	0.9999	0.00160	2.576e-06	3.348e-05
Random forest	0.9999	0.00095	9.144e-7	2.8309e-5
Support Vector	0.9850	0.02573	0.000666	0.01997
Kneighbors	0.9997	0.003614	1.3064e-5	0.002179

Table 1: Comparison table of different ML models

The given table 1 is the summary of the different metrics used to predict the model accuracy. In the given table we can see that the highest accuracy for the decision tree and random forest regression mode. similarly, the RMSE value is least for the Random Forest regression mode, the other metrics that is MSE and MAE is also least for the Random Forest Regression model, which shows the Random Forest is the best method for the air quality prediction.

V. CONCLUSION:

In conclusion, our investigation into Indian air quality analysis and prediction using machine learning highlights the crucial need for innovative solutions to meet the nation's rising environmental concerns.

As the industrial growth in India picks up speed, drastic measures are required to combat the rising levels of pollutants like carbon dioxide and particulate matter. In the quest for efficient air quality monitoring and forecasting, our novel machine learning model—which boasts an astounding 99% accuracy rate—offers promise. Instead of

focusing only on algorithms and statistics, this study examines the air we breathe and how it affects the health and well-being of the Indian population.

We not only increase our understanding of pollution trends by utilizing the power of machine learning, but we also empower policymakers and communities to make informed decisions. The combination of historical data and a wide range of machine learning algorithms ushers in a new era of environmental stewardship. As we aim for a cleaner, healthier future, this research represents a crucial step forward in our collaborative effort to safeguard the air we breathe and, as a result, the lives of those who live in India.

VI. REFRENCES

[1] Gnana Soundari, A., Gnana Jeslin, J., & Akshaya, A. C. (2019). Indian air quality prediction and analysis using machine learning. *International Journal of Applied Engineering Research, 14*(11) (Special Issue), ISSN 0973-4562

[2] B. D. Parameshachari, G. M. Siddesh, V. Sridhar, M. Latha, K. N. A. Sattar and G. Manjula., "Prediction and Analysis of Air Quality Index using Machine Learning Algorithms," 2022 IEEE International Conference on Data Science and Information System (ICDSIS), Hassan, India, 2022, pp. 1-5, doi: 10.1109/ICDSIS55133.2022.9915802

[3] A. Akanksha, N. Maurya, M. Jain and S. Arya, "Prediction And Analysis of Air Pollution Using Machine Learning Algorithms," 2023 3rd International Conference on Intelligent Technologies (CONIT), Hubli, India, 2023, pp. 1-6, doi: 10.1109/CONIT59222.2023.10205615.

[4] V. Hable-Khandekar and P. Srinath, "Machine Learning Techniques for Air Quality Forecasting and Study on Real-Time Air Quality Monitoring," 2017 International Conference on Computing, Communication, Control and Automation (ICCUBEA), Pune, India, 2017, pp. 1-6, doi: 10.1109/ICCUBEA.2017.8463746.

[5] Madan, Tanisha & Sagar, Shrddha & Virmani, Dr. Deepali. (2020). Air Quality Prediction using Machine Learning Algorithms –A Review. 140-145. 10.1109/ICACCCN51052.2020.9362912.

[6] K. M. O. V. K. Kekulanadara, B. T. G. S. Kumara and B. Kuhaneswaran, "Machine Learning Approach for Predicting Air Quality Index," 2021 International Conference on Decision Aid Sciences and Application (DASA), Sakheer, Bahrain, 2021, pp. 622-626, doi: 10.1109/DASA53625.2021.9682221.

[7] C. Shi, Y. Wang, Y. Wan and S. Wu, "Air Quality Prediction Based on Machine Learning," 2022 International Conference on Machine Learning and Knowledge Engineering (MLKE), Guilin, China, 2022, pp. 1-5, doi: 10.1109/MLKE55170.2022.00008.

[8] Kaur, Gaganjot & Gao, Jerry & Chiao, Sen & Lu, Shengqiang & Xie, Gang. (2018). Air Quality Prediction: Big Data and Machine Learning Approaches. International Journal of Environmental Science and Development. 9. 8-16. 10.18178/ijesd.2018.9.1.1066.

[9] Kumar, K., Pande, B.P. Air pollution prediction with machine learning: a case study of Indian cities. Int. J. Environ. Sci. Technol. 20, 5333–5348 (2023). https://doi.org/10.1007/s13762-022-04241-5

[10] Bekkar, A., Hssina, B., Douzi, S. et al. Air-pollution prediction in smart city, deep learning approach. J Big Data 8, 161 (2021). https://doi.org/10.1186/s40537-021-00548-1

[11] Kumar Singh, Rohit and Raghav, Shekhar and Maini, Tarun and Kumar Singh, Murari and Arquam, Md., Air Quality Prediction using Machine Learning (July 14, 2022). Proceedings of the Advancement in Electronics & Communication Engineering 2022

[12] Nehete, R., & Patil, D. D. (2021). Air quality prediction using machine learning. International Journal of Current Research and Technology (IJCRT), 9(6), ISSN: 2320-2882.