International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

# Air Quality Prediction Using Machine Learning

*Ajaykumar C*

Rani Channamma University, Belagavi

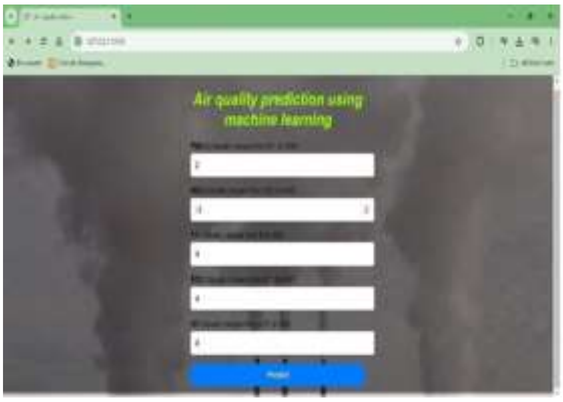
**ABSTRACT**

The prediction of air quality is essential for monitoring the environment and managing public health. This research investigates using machine learning algorithms to forecast air quality indices (AQI) based on different atmospheric factors. By utilizing past air quality data and meteorological variables, we created models to predict AQI levels. Our method includes preparing the data, selecting features, training models, and assessing them using various machine learning algorithms such as Linear Regression and Random Forests. The findings illustrate the potential of machine learning methods in accurately forecasting air quality, offering valuable insights for proactive air quality control

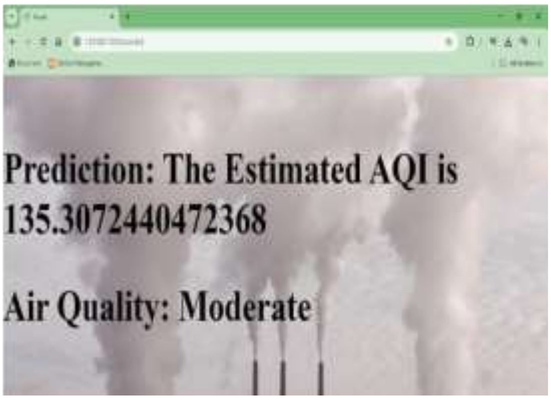
**KEYWORDS:** Air Quality Prediction, Machine Learning, Air Quality Index, Environmental Monitoring, Predictive Models, Linear Regression, Random Forests, Neural Networks.

## 1. INTRODUCTION

Air is crucial for sustaining life, as the human body relies on inhaling oxygen for cellular functions. While humans can survive without water for days, they can only last a few minutes without air. In addition to sustaining life, air plays a critical role in regulating the Earth’s temperature and supporting the water cycle. However, the quality of the air we breathe directly impacts our health. Poor air quality, caused by pollution, can lead to respiratory illnesses such as bronchitis, asthma, pneumonia, and even lung cancer. It is estimated that air pollution claims about 7 million lives annually worldwide. Air pollution also contributes to global warming by trapping heat in the atmosphere, leading to rising temperatures, sea level increases, and the spread of heat-related illnesses and infectious diseases. The **Air Quality Index (AQI)** is a valuable measure used to evaluate air quality in a specific area. The AQI scale ranges from 0 to 500, with higher values indicating worse air quality and greater health risks. An AQI of 50 or below indicates good air quality, while levels above 300 represent hazardous conditions. For easy interpretation, the AQI is categorized into six color-coded groups: Green (0-50), Yellow (51-100), Orange (101-150), Red (151- 200), Purple (201-300), and Maroon (300+). Monitoring and calculating AQI is crucial to provide timely warnings to the public about potentially harmful air conditions. It helps individuals, especially vulnerable groups, take protective measures when air quality is poor. AQI highlights the immediate health risks from breathing polluted air over short periods, typically hours to days. Machine learning offers a powerful set of tools for predicting air quality. Machine learning models are trained on data to identify patterns and make informed predictions. There are three primary types of machine learning: supervised, unsupervised, and semi-supervised. Supervised learning, which was utilized in this project, relies on labeled datasets where each input has a known output. This approach tends to provide more accurate results compared to unsupervised methods, especially when there is abundant labeled data available.



Air quality prediction using machine learning



Prediction: The Estimated AQI is 135.3072440472368

Air Quality: Moderate

**Figure 1: dash bord  Figure 2: Quality of air**

## 2. LITERATURE SURVEY:

Throughout this research, a thorough investigation is proposed for using deep learning architectures to model contamination data [1].A combined model employing Artificial Neural Networks and Kriging was utilized to assess the air pollutant levels at different locations in Mumbai [2].MATLAB was used for the Artificial Neural Network (ANN) and R for Kriging to predict future pollution based on fundamental parameters, employing Linear regression and Multilayer Perceptron Protocol (ANN) [3].Logistic regression is utilized to identify polluted or unpolluted data samples, while Autoregression is used to forecast future PM2.5 values based on previous readings [4].Data from Shenzhen, China was used to enhance and visualize air quality maps, employing algorithms such as Artificial Neural Network (ANN), Genetic Algorithm ANN Model, Random forest, decision tree, and Deep belief network, and discussing the various strengths and weaknesses of the model [5].A Stacked Auto-encoders model is employed to learn and train data, proposing a deep learning approach for predicting pollution in South Korea [6].The approach involved predicting the air quality standard for the next 48 hours at each monitoring station, considering air quality data, meteorology data, and weather outlook data [7].A model was developed to predict the air quality index, utilizing historical data from previous years and predicting over a specific upcoming year, as a Gradient decent boosted multi-variable regression problem [8].In a recent study [9], a more sophisticated model was utilized to forecast hourly air pollution patterns by using previous days' meteorological data, treating the 24- hour prognosis as a multi-task learning (MTL) problem. The proposed prediction model resulted in a 57%, 47%, 47%, and 94% reduction in error percentage, making it the most reliable algorithm for contamination prediction [10].Another study [11] focused on understanding the behavior of PM2.5 emissions in the 50 most polluted capital cities, comparing data before and after the implementation of quarantine measures. Various models, such as SVM and ANN, were developed using meteorological and pollutant parameters from 2016-18. The performance of these models in predicting PM2.5 levels was thoroughly evaluated and discussed [12].The air quality data cited in this paper [13] is sourced from the India air quality monitoring and analysis platform, encompassing daily averages of PM2.5, PM10, O3, CO, SO2, NO2 concentrations, and the air quality index (AQI).In a different study [14], models for hourly air quality forecasting in California were built using support vector regression (SVR), a powerful machine learning approach. These models aimed to forecast pollutant levels and accurately determine the AQI.A case study on Delhi and Houston focused on forecasting AQI using regression models. Different models, including SVR and linear models like multiple linear regression with gradient descent variations, were implemented. Among these, SVR demonstrated superior performance in various quality measures [15].

## 3. METHODOLOGY:

The technique for predicting air satisfactory using gadget learning entails a series of well-based steps that embody records series, preprocessing, feature choice, version choice, education, evaluation, and deployment. each step the critical on making sure this the prediction version is each accurate and reliable. beneath, we define the technique in element.

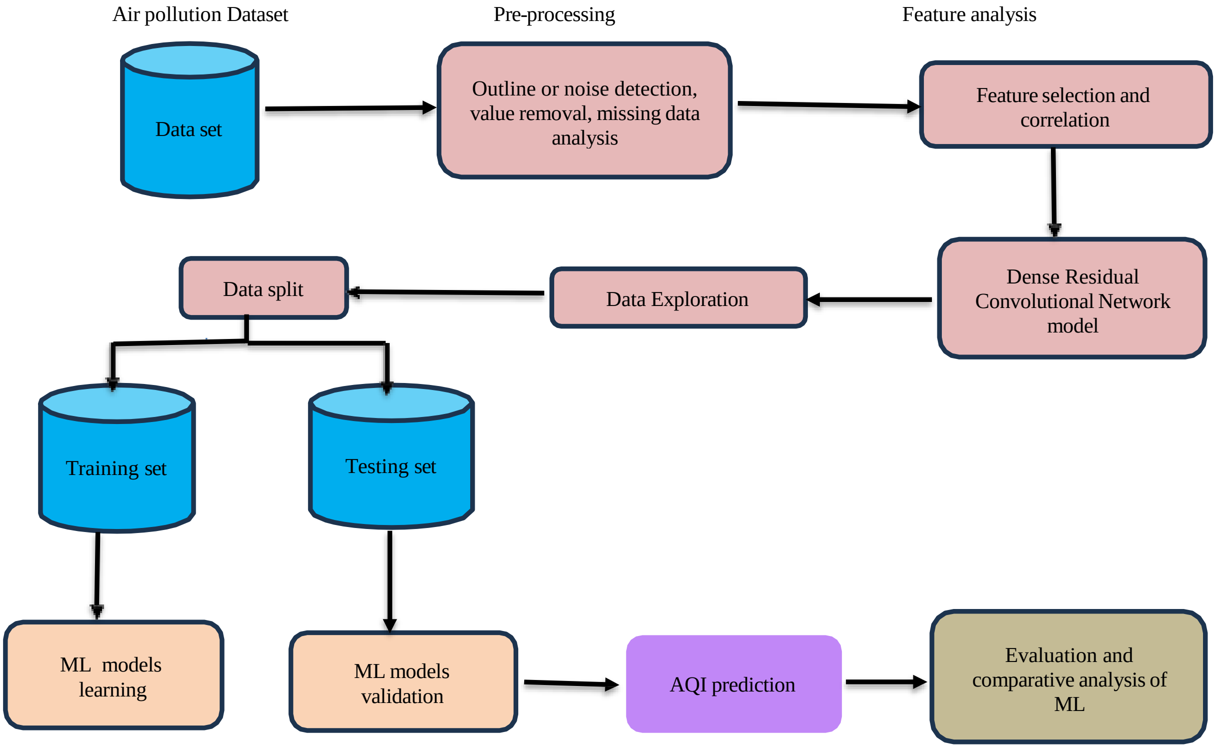

Air pollution Dataset
Pre-processing
Feature analysis
Data set
Outline or noise detection, value removal, missing data analysis
Feature selection and correlation
Dense Residual Convolutional Network model
Data Exploration
Data split
Training set
Testing set
ML models learning
ML models validation
AQI prediction
Evaluation and comparative analysis of ML

**Figure4.1: System Architecture Air Quality Prediction Using machine learning**

.**Air pollutants Dataset:**

- **Dataset:** The dataset used in this workflow comes from the includes air pollutants information particular to India. This dataset could be used to educate and compare the machine mastering model.

**Pre-processing**

- **Outlier or Noise Detection, cost removal, lacking facts analysis:**

The pre-processing step includes cleaning the statistics through detecting and managing outliers or noise, getting rid of (no longer various) values, and reading any missing records. This ensures that the dataset is correct and complete before function evaluation.

**Function evaluation:**

- **feature selection and Correlation:** in this step, the relevant capabilities are selected primarily based on their correlation with the goal variable (AQI). This allows in figuring out the maximum vital capabilities in an effort to affect the prediction version.
- **Dense Residual Convolutional community model**:

A Dense Residual Convolutional network blended with a used to seize each spatial and temporal dependencies in the statistics. This machine getting to know version is designed to system the features selected in the preceding step and generate predictions.

**Records Exploration**

- earlier than splitting the dataset, information exploration is performed to apprehend the distribution, developments, and styles inside the records. This step helps in making knowledgeable choices approximately version choice and schooling.

**Data Split**

- **The Dataset Is Cut UpIn to Elements:** a education Set to trying out the Set. The education set used in train that version, whilst the checking out set that is used validate its performance. ML Modals mastering & ML Validation
- **ML studying:** The ML getting to know model is skilled at the education set. in the course of this procedure, the version learns the underlying styles and relationships among the functions and AQI.
- **ML models Validation:** After schooling, the model is established at the testing set. This step is critical to ensure that is model using generalizes well to new, unseen facts.

**AQI Prediction**

- as soon as confirmed, the model the used on the predict the AQI primarily based on new information inputs. assessment and Comparative evaluation of ML
- The very last step entails comparing the overall performance of the system getting to know version and evaluating it with different fashions or techniques. This analysis facilitates in understanding the strengths and weaknesses of the model and possibly refining it similarly.

**TABLE:4.1 AQI BASICS FOR OZONE INDEX AND PARTICLE POLLUTION**

| Air Quality Index (AQI) Values | Levels of Health Concern | Colors |
|---|---|---|
| 0 to 50 | Good | Green |
| 51 to 100 | Satisfactory | Yellow |
| 101 to 150 | Moderate | Orange |
| 151 to 200 | Poor | Purple |
| 201 to 300 | Severe | Red |

- "correct" AQI is zero to 50. Air fine is taken into consideration first-rate, and air pollution poses very little danger.
- " best " AQI is 51 to one hundred. Air best is suitable; but, for some pollution there can be a slight fitness concern for a very small quantity of human beings. as an instance, those who are surprisingly sensitive to ozone may additionally experience respiratory signs and symptoms.
- “mild” air exceptional is an AQI among one hundred and one-one hundred fifty and is represented by way of an orange color. At this category individuals of touchy groups can also enjoy fitness effects. anybody with coronary heart or lung sickness, older adults, everybody pregnant, and children (such as teens), are all considered a touchy institution. the majority is less possibly to be affected at this class, but all and sundry have to avoid out of doors activities that reason heavy respiration, like operating out, going for walks, or gambling sports activities. touchy businesses have to limit how long they spend outside.

- "poor" air high-quality" is an AQI among 151-two hundred and is represented through a pink shade. At this class some individuals of most of the people may additionally enjoy fitness results; participants of touchy groups may additionally enjoy greater serious fitness effects. absolutely everyone must restriction how tons time is spent outdoor, and sensitive companies have to avoid going out of doors as an awful lot as feasible..
- "excessive" air excellent" is an AQI among 201-300 and is represented by a purple colour. At this class, the danger of fitness results is elevated for all people. while air satisfactory is this terrible outside, all of us have to avoid going outdoor as tons as feasible. carrying correctly geared up N95 masks is recommended while outdoor.

**HOW IS THE AIR QUALITY INDEX (AQI) CALCULATED**

Ip= $\underline{\text{IH1-ILo (CP – BP Lo}}$) +ILO. BPHI - BPLO

- Corr Where Ip= the index For pollutant P
- Cp= the truncated concentration of pollutant P
- BPHI= the concentration breakpoint that is greater then or equal to Cp
- BPLO= the concentration breakpoint that is less then or equal to Cp
- IHI= the AQI value corresponding to BPHI
- ILO= the AQI values corresponding to BPLO

he PM2.five AQI is an index developed by way of diverse governmental companies inclusive of the USEPA, or China's CNEMC, to deliver how polluted the air is with recognize to PM2.5. The computational formulae used for AQI calculation varies depending on the regulatory jurisdiction. For the motive of our dialogue, we can use the computations utilized by the unitedstates-EPA.

As we will see, the AQI isn't continually an amazing predictor for destiny air fine. in contrast to direct PM measurements, AQI is a unitless number that varies from 0 to greater than 500. PM2.five AQI is a midnight-to-nighttime 24-hour cost based on 1-hour measured values. The PM2.5 AQI is computed from the following components where Ip = AQI. See fig1.2. 24 1-hour measured PM values from nighttime to middle of the night are had to compute the Air fine Index "AQI." on this computation, it is vital to understand the breakpoints between the AQI categories

3.2 facts series

the primary and main step in the methodology is the gathering of facts. For air exceptional prediction, facts can be sourced from numerous governmental companies, environmental tracking stations, and open-source databases. commonplace datasets include facts on pollutants consisting of PM2.five, PM10, NO2, SO2, CO, O3, and meteorological information inclusive of temperature, humidity, wind velocity, and course. ancient data on these parameters over several years is vital to recognize tendencies and patterns.

in addition to these traditional sources, information also can be collected from IoT-based totally sensors deployed in particular areas for actual-time tracking. the mixing of those information factors enables in building a complete dataset that can better mirror the dynamic nature of air excellent.

3.2 records Preprocessing

as soon as the records is collected, it undergoes a rigorous preprocessing phase. This step is critical due to the fact the first-class of the information immediately influences the overall performance of the system-studying version. The preprocessing steps consist of:

- information cleaning: getting rid of noise, managing missing values, and correcting inaccurate entries. for example, lacking values can be imputed using strategies which include mean, median, or mode imputation, or more advanced methods like k-Nearest neighbors (KNN) imputation.
- Normalization and Scaling: for the reason that dataset may additionally include variables of various devices, normalization or scaling is done to bring all of the functions onto a commonplace scale. this is especially vital for algorithms like aid Vector Machines (SVM) and ok-nearest pals, which might be sensitive to the magnitude of the enter information.
- managing Outliers: Outliers within the facts can skew the model's predictions. strategies which include Z-score analysis or the IQR approach are used to perceive and take care of those outliers.
- function Engineering: New capabilities that can improve the version's performance are created. as an instance, combining temperature and humidity to create a brand new feature like "warmth Index" can from time to time cause higher predictions. Temporal features, together with the time of day or season, also are introduced to capture the cyclical nature of air fine modifications.

three.3 characteristic selection

feature selection is the manner of figuring out the most huge variables that make contributions to predicting air great. This step is crucial because it reduces the dimensionality of the facts, improving model performance and lowering computation time.

## 4. HARDWARE AND SOFTWARE REQUIREMENTS

**Hardware Requirements**

|  |  |
|---|---|
| Processor: | Pentium Dual Core 2.00GHZ |
| Output Devices: | Monitor (LCD) |
| Input Devices: | Keyboard |
| Hard Disk: | 1TB |
| RAM: | 8GB Or Above |

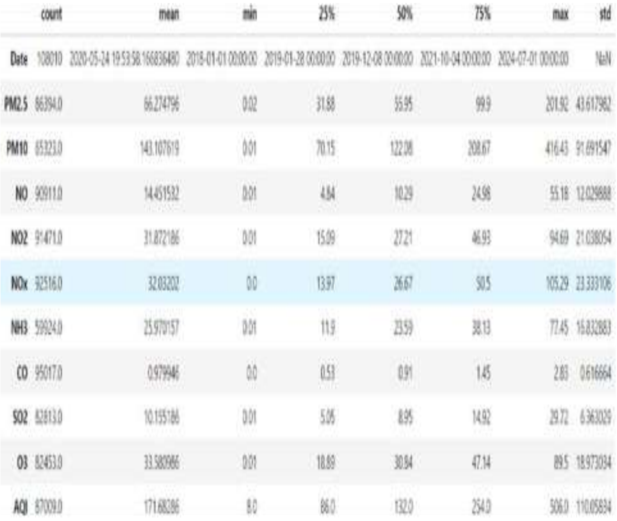
**Software Requirements**

|  |  |
|---|---|
| Scripting language: | Python Programming |
| Scripting Tool: | Anaconda Navigator (Jupyter Notebook) & Visual studio Code |
| Operating system: | Microsoft Windows 10 or 11 |
| Dataset: | Excal.csv File |
| Machine Learning Packages: | NumPy, Pandas, Matplotlib, Seaborn Packages GUI: Python frame works Flask |

### *4.1 Linear Regression Model:*

It is a fundamental statistical and machine learning technique, which is used to model the relationship between one dependent variable (target) and one or more independent variables (features). The goal is to find the linear relationship between the two and it explains how the dependent variable changes with the independent variables

## 5. RESULTS AND SCREENSHOTS:



|  | count | mean | min | 25% | 50% | 75% | max | std |
|---|---|---|---|---|---|---|---|---|
| Date | 108010 | 2020-05-24 19:53:58.166836480 | 2018-01-01 00:00:00 | 2019-01-28 00:00:00 | 2019-12-08 00:00:00 | 2021-10-04 00:00:00 | 2024-07-01 00:00:00 | NaN |
| PM2.5 | 86394.0 | 66.274796 | 0.02 | 31.88 | 55.95 | 99.9 | 201.92 | 43.617982 |
| PM10 | 65323.0 | 143.107619 | 0.01 | 70.15 | 122.08 | 208.67 | 416.43 | 91.691547 |
| NO | 90911.0 | 14.451532 | 0.01 | 4.84 | 10.29 | 24.98 | 55.18 | 12.029888 |
| NO2 | 91471.0 | 31.872186 | 0.01 | 15.09 | 27.21 | 46.93 | 94.69 | 21.038054 |
| NOx | 92516.0 | 32.03202 | 0.0 | 13.97 | 26.67 | 50.5 | 105.29 | 23.333106 |
| NH3 | 59924.0 | 25.970157 | 0.01 | 11.9 | 23.59 | 38.13 | 77.45 | 15.832883 |
| CO | 95017.0 | 0.979946 | 0.0 | 0.53 | 0.91 | 1.45 | 2.83 | 0.616664 |
| SO2 | 82813.0 | 10.155186 | 0.01 | 5.05 | 8.95 | 14.92 | 29.72 | 6.363029 |
| O3 | 82453.0 | 33.580986 | 0.01 | 18.89 | 30.84 | 47.14 | 89.5 | 18.973034 |
| AQI | 87009.0 | 171.68286 | 8.0 | 86.0 | 132.0 | 254.0 | 506.0 | 110.05834 |

Figure 4: Dataset of AQI

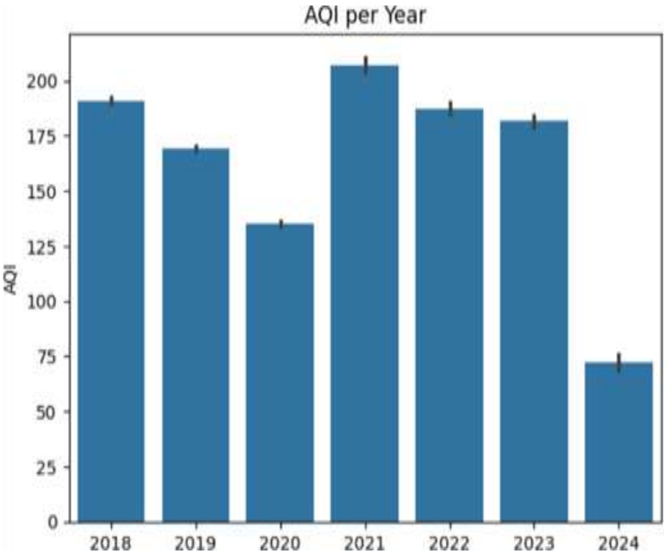


Figure 3: AQI per year

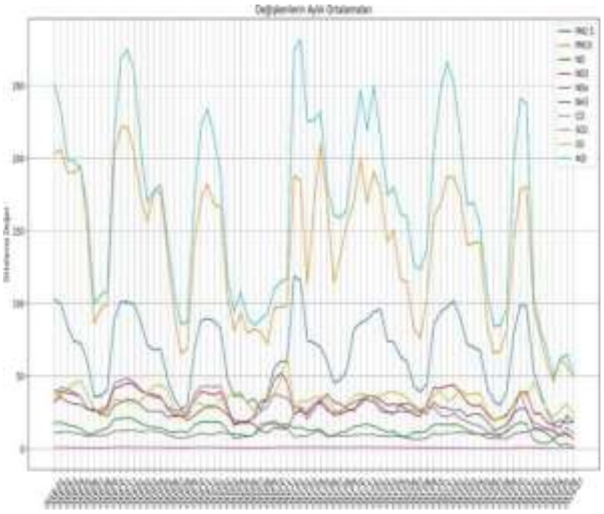


Figure 6: ortalama degree AQI

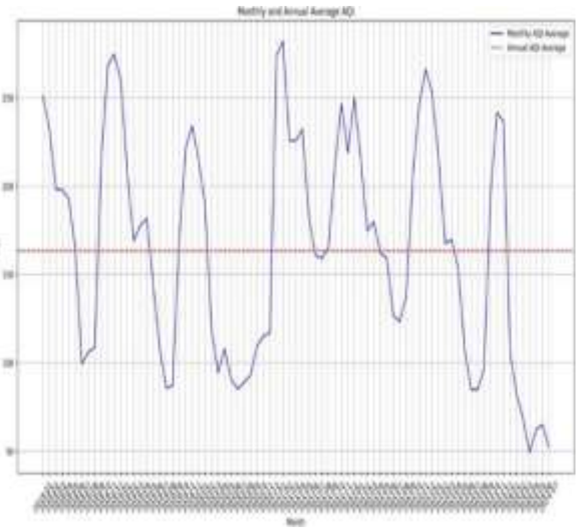


Figure 7: Monthly and Annual Average AQI

## 6. CONCLUSION

The project successfully developed and implemented a machine learning model to predict air quality based on key pollutants, including PM2.5, NO, SO, and O₃. By leveraging historical data, the model provided accurate predictions of air quality indices, demonstrating the critical role these pollutants play in determining air quality levels.

**Applications of the Major Project**

- The model achieved a high level of accuracy, particularly in predicting PM2.5 levels, which are a significant indicator of air quality. The use of features such as NO, SO, and O₃ further improved the model's predictive capability.
- Importance of Among the pollutants, PM2.5 was found to be the most significant contributor to poor air quality. This aligns with existing research highlighting the dangers of fine particulate matter to human health.
- Impact of NO, SO, and O₃: Nitric Oxide (NO), Sulfur Dioxide (SO), and Ozone (O₃) were also identified as important factors in air quality prediction. Their interactions with PM2.5 and each other suggest that air quality is a multifactorial issue, requiring comprehensive monitoring and regulation.
- Model Applications The model can be utilized by environmental agencies for real-time air quality monitoring and forecasting. This would enable timely public health interventions and policy adjustments, particularly in urban areas where pollutant levels fluctuate frequently.
- Despite the model's success, challenges such as data quality, missing values, and the need for more granular data were encountered. Future work could focus on incorporating additional features like meteorological data (temperature, humidity, wind speed) to further enhance predictive accuracy.

**ACKNOWLEDGEMENT**

I am grateful to Dr.Mallamma V.Reddy Assistant Professor Department of Computer Science, Rani Channamma University, Belagavi for his valuable guidance for completion of this work

**REFERENCES:**

1. Ak Yasin Ayturan, Zeynep CansuAyturan. Air Pollution Modelling With Learning: A Review. September 2018.
2. Suhasini V. Kottur , Dr. S. S. Mantha. An Integrated Model Using Artificial Neural Network (Ann) And Kriging for Forecasting Air Pollutants Using Meteorological Data. International Journal of Advanced Research in Computer and Communication Engineering ISSN (Online): 2278-1021 ISSN (Print) :2319-5940 Vol. 4, Issue 1, January 2015.
3. Ruchi Raturi, Dr. J.R. Prasad. Recognition Of Future Air Quality Index Using Artificial Neural Network. International Research Journal of Engineering and Technology (IRJET) .e-ISSN: 2395- 0056p. ISSN: 2395-0072 Volume: 05 Issue: 03 Mar-2018.
4. Aditya C R, Chandana R Deshmukh, Nayana D K, Praveen Gandhi Vidyavastu. Detection and Prediction of Air Pollution using Machine Learning Models. International Journal of Engineering Trends and Technology (IJETT) – volume 59 Issue 4 – May 2018.
5. Gaganjot Kaur Kang, Jerry Zeyu Gao, Sen Chiao, Shengqiang Lu, and Gang Xie Air Quality Prediction: Big Data and Machine Learning Approaches. International Journal of Environmental Science and Development, Vol. 9, No. 1, January 2018.

6. ThanongsakXayasouk, Hwamin Lee. Air Pollution Prediction System Using Deep Learning. 2018.
7. Xiuwen Yi, JunboZhang ,Zhaoyuan Wang, Tianrui Li ,Yu Zheng. Deep Distributed Fusion Network for Air Quality Prediction. KDD '18: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. July 2018.
8. A. GnanaSoundari, J. Gnana Jeslin, Akshaya A.C. Indian Air Quality Prediction And Analysis Using Machine Learning. International Journal of Applied Engineering Research ISSN 0973-4562 Volume 14, Number 11, 2019.
9. DixianZhu ,Changjie Cai , Tianbao Yang and Xun Zhou. A Machine Learning Approach for Air Quality Prediction: Model Regularization and Optimization. 24 February 2018.
10. Mahmoud Reza Delavar, Amin Gholami, Gholam Reza Shiran, Yousef Rashidi, Gholam Reza Nakhaeizadeh, Kurt Fedra and SmaeilHatefi Afshar. A Novel Method for Improving Air Pollution Prediction Based on Machine Learning Approaches: A Case Study Applied to the Capital City of Tehran.
11. Daniella Rodríguez-Urrego, Leonardo Rodríguez-Urrego. Air quality during the COVID-19: PM2.5 analysis in the 50 most polluted capital cities in the world. Environmental Pollution. 266(2020) 115042.
12. Adil Masood, Kafeel Ahmed. A model for particulate matter (PM2.5) prediction for Delhi based on machine learning approaches. Procedia Computer Science 167 (2020) 2101–2110. PJAEE, 17(7) (2020) 7003
13. Mauro Castelli ,Fabiana Martins Clemente,AlesPopovic,Sara Silva and Leonardo Vanneschi. A Machine Learning Approach to Predict Air Quality in California. Hindawi Complexity, Volume 2020, Article ID 8049504, 23 pages.
14. Doreswamy, Harishkumar K S1, Yogesh KM, Ibrahim Gad. Forecasting Air Pollution Particulate Matter (PM2.5) Using Machine Learning Regression Models. Procedia Computer Science 171 (2020) 2057–2066
15. S. S. Ganesh, S. H. Modali, S. R. Palreddy, and P. Arulmozhivarman, “Forecasting air quality index using regression models: A case study on delhi and houston,” in 2017 International Conference on Trends in Electronics and Informatics (ICEI), 2017, pp. 248–254.