

**BRACT’s**

**Vishwakarma Institute of Information Technology**

**RESEARCH PAPER ON AQI Prediction using ML Algorithms**

**Department of Computer Science and Engineering (Artificial Intelligence)**

|  |  |  |  |
| --- | --- | --- | --- |
| Roll no. | PRN no. | Name | Email |
| 281045 | 22210675 | Sanchit Motghare | sanchit.22210675@viit.ac.in |
| 281033 | 22210636 | Samruddhi Kathale | samruddhi.22210636@viit.ac.in |
| 281044 | 22210898 | Abhishek More | abhishek.22210898@viit.ac.in |
| 281012 | 22211524 | Swapnil Darphale | swapnil.22211524@viit.ac.in |

Under Guidance of

Mr. Amol Patil

**AQI Prediction using ML Algorithms (ARIMA and RFR)**

### **Abstract**

The escalating concerns surrounding air pollution have underscored the urgent need for accurate forecasting of Air Quality Index (AQI) levels. This calls for advanced predictive models to analyze its adverse impacts. This research endeavors to develop a robust predictive framework utilizing Random Forest Regression (RFR) and Autoregressive Integrated Moving Average (ARIMA) models to forecast AQI values in urban environments. We leverage a comprehensive dataset spanning key cities in India from 2015 to 2020 containing crucial information about several pollutant concentrations and AQI readings across 25 cities including major cities like Delhi, Ahmedabad, Mumbai, Chennai, Amravati etc. Our methodology integrates rigorous data pre-processing techniques to ensure data integrity and model suitability. The RFR model harnesses ensemble learning to capture complex relationships between pollutants and AQI, while ARIMA excels in modeling temporal dependencies and seasonality patterns. Through meticulous model training and evaluation, we aim to provide actionable insights for environmental management and public health initiatives. Our research presents promising findings from both Random Forest Regressor (RFR) and Autoregressive Integrated Moving Average (ARIMA) models for forecasting Air Quality Index (AQI) values. RFR yielded a low Mean Squared Error (MSE) of 1689.24 and a high R-squared (R2) value of 0.908, signifying strong predictive accuracy and explanatory power. Similarly, the ARIMA model demonstrated favorable performance with a Root Mean Squared Error (RMSE) of approximately 25.12 units and a Mean Absolute Percentage Error (MAPE) of around 16.52%, underscoring its efficacy in capturing AQI trends accurately.

### **Keywords**

Air Quality Index (AQI), Random Forest Regression (RFR), Autoregressive Integrated Moving Average (ARIMA), pollutant concentrations, data pre-processing, ensemble learning

### **Introduction**

Air pollution has become a significant concern globally because of industrialization, urbanization, and a variety of other contributing elements. The Air Quality Index (AQI) is a crucial measure indicating the extent of atmospheric pollution, with its value intricately tied to the presence of various pollutants present in the atmosphere. Specifically, six key pollutants, namely PM2.5, PM10, NO2, SO2, CO, and O3, have a notable impact on air quality levels. Prolonged exposure to air pollution has been associated with a spectrum of health issues, ranging from respiratory to cardiovascular ailments. Despite its multifaceted nature, air pollution prediction presents a formidable challenge, owing to the non-constant and nonlinear characteristics of pollutant concentrations.

Scholars have examined a range of prediction models in recent times, which can be broadly categorized into statistical, machine learning, and deep learning paradigms. For the purpose of explaining causation, statistical models like the autoregressive (AR) model and autoregressive integrated moving average (ARIMA) model depend on presumptions about the arrangement of data and parameter inference. In contrast, machine learning models prioritize model prediction above parameter estimation by using big datasets to anticipate future events. Prominent research works have used machine learning methods to predict air pollution laminar combustion rates and comprehend the effects of indoor air quality variations.

The proliferation of industrial and daily activities has intensified air pollution concerns, particularly in burgeoning urban centers like China and India. Exposure to outdoor pollutants, including ozone, nitrogen oxides, volatile organic compounds, and particulate matter, has been linked to adverse health effects and increased mortality rates. Notably, PM2.5 emerges as a critical pollutant, with its concentrations calculated using logistic regression and autoregression methods. The prediction of pollution levels often involves hourly data analysis, necessitating active air quality monitoring across urban environments. Although current air quality assessment relies heavily on static monitoring stations, advancements in predictive algorithms offer promising avenues for real-time pollution monitoring and mitigation strategies.

Air pollution’s deleterious effects extend beyond human health, impacting ecological balance and agricultural productivity. Mitigating these impacts requires a multifaceted approach, combining advanced modeling techniques with robust monitoring infrastructure. By harnessing machine learning and statistical modeling, researchers aim to develop accurate and actionable insights into air quality dynamics, thereby contributing to sustainable environmental stewardship and public health initiatives.

The objective of this research project is to create a resilient and precise predictive model for forecasting the Air Quality Index (AQI) in urban settings, with an emphasis on integrating multiple machine learning techniques for enhanced prediction capabilities. Specifically, this project employs Random Forest Regression (RFR) and Autoregressive Integrated Moving Average (ARIMA) models to forecast AQI values based on various input parameters, including date and concentrations of key air pollutants.

The project is divided into two main components: data pre-processing and model development. In the data pre-processing phase, the collected dataset, comprising historical air quality data and pollutant concentrations across different urban areas, undergoes extensive pre-processing steps to ensure data quality and uniformity. This involves replacing null values with different techniques like either taking a mean, or using some traditional methods like b-fill etc. Additionally, the dataset is structured to include relevant input features such as date and pollutant concentrations.

For model development, the project leverages the capabilities of two distinct machine learning techniques: Random Forest Regression (RFR) and Autoregressive Integrated Moving Average (ARIMA). The RFR model is chosen for its ability to handle multiple input variables and capture nonlinear relationships, making it well-suited for predicting AQI values based on various pollutant concentrations and temporal factors. On the other hand, the ARIMA model is employed for its efficacy in apprehending temporal dependencies and seasonal patterns in time series data, thus bolstering the precision of AQI forecasts.

### **Literature Survey**

### In this work [1], we first checked the equating betwixt a number of air signs, containing the AQI, PM2.5 concentrations, total NOx (nitrogen oxides) concentrations, thus. Second, we secondhand haphazard wood regression (RFR) to devise guess models. Lastly, we secondhand equivalence cooperative r and coefficient of perseverance (RSQUARE) to judge the accomplishment of the reversion models.

### The air condition index is forecasted, and pollutant and piece levels are determined, utilizing the familiar machine intelligence technique SVR [2]. The results show that SVR accompanying the RBF essence can correctly forecast at fixed intervals AQI for united states of america of California, in addition to hourly concentrations of contaminants to a degree colorless odorless toxic gas, sulphur dioxide, nitrogen dioxide, close to the ground air, and particulate matter 2.5. Using the US Environmental Protection Agency's dataset, the earlier hidden confirmation dossier was top-secret into six AQI types with 94.1 allotment veracity.

### The AQI forecasting exploits machine intelligence methods like opportunity succession study and LR. MLR and directed machine intelligence methods were employed to forecast the AQI. A range of mathematical versification were working to judge the influence. Second, the ARIMA time order model was working to project the AQI from now on. It was found that two together models were completely effective and correct at concluding the AQI [3].

### Artificial affecting animate nerve organs networks and the Kriging form were working in an integrated model to estimate the amount of air contamination at differing neighborhoods in Mumbai and Navi Mumbai. The extreme R principles designated that the required scope of arrangement betwixt the noticed and wonted values had happened attained. ANN acted better than fundamental reversion models in conditions of forecast and R value [4].

### AQI biographer aggregation prophecy utilizing limits such as PM2.5, PM10, SO2, and NO2. To sum up, the chance thicket reversion treasure caused highest in rank results out of the conclusion timber reversion, uninterrupted reversion, SVR, and RFR algorithms. The test data accompanying defeater in competition veracity of 0.99985, the slightest mean square mistake of 0.00013, and the mean absolute mistake of 0.00373 [5].

### Using dossier from the former old age and jutting over a named future year as a slope lowering embellished the multivariable reversion question, the AQI was forecasted. They surpassed traditional reversion arrangements by improving the model's influence by utilizing forecasted question cost estimations. In order to determine the order of option established by means of what carefully the alternatives approximated the ideal solution, they further working the AHP MCDM method [6].

### This paper [7] compares Random Forest, Linear Regression, and Decision Tree reversion for air character guess, assessing versification like Mean Absolute Error and R2 score. This study compares reversion models for air characteristic forecast in smart capitals, finding Decision Tree reversion superior. It embellishes veracity accompanying Exploratory Data Analysis and SMOTER method, and stresses cloud computing benefits for effectiveness. Future work involves surveying different machine intelligence approaches and integrating of or in the atmosphere dossier for revised guess veracity and air status management.

### This [8] research presents an advanced machine intelligence model joining Grey Wolf Optimization definitely Tree regression to call Air Quality Index (AQI) in big Indian downtowns. They have secondhand dataset from Kaggle , metropolises healthy Delhi, Kolkata, Chennai, Bangalore, and Visakhapatnam are analysed. The model's performance, judged through versification like R-Square, RMSE, MSE, MAE, and veracity, outperforms usual machine intelligence algorithms. The study suggests potential continuation accompanying deep education models for further improved forecasting veracity in air quality listening.

### Over the course of a old age (August 2009–August 2010), the ANN treasure forecasted the at fixed intervals test pollutants aggregation levels in addition to the AQI and AQHI for Ahvaz, Iran. This study demonstrated that, in consideration of thwart negative energy impacts, the ANN can be used to forecast air value in centers like Ahvaz. They decided that a fake interconnected system might be secondhand by city air value managers to determine the spatiotemporal description of contaminants and air kind versification [9]. Tree-located ensemble knowledge models were conceived to examine the city air value of Lucknow, India, over a five-old age ending, utilizing of or in the atmosphere and air quality records. To find the beginnings of air contamination, PCA was working. The use of pushing and bagging techniques admitted the DTF and DTB models to beat the SVM in categorization and reversion. They urged ensemble models for ruling the ambient air character in places were intelligent to correctly forecast it.

### [10] The AQI was correctly predicted utilizing an ML-located order established facts assembled from meteorological stations and preservation of natural resources. The forecast approach create use of an embellished interconnected system that is to say expressly planned for a short time-order indicator: a new nonlinear autoregressive neural network (ARNN) accompanying an exogenic recommendation model. The example was used to a research project that complicated several London-extent weathers listening stations [11].

### In this work, the following algorithms—directed knowledge, support heading machines, and affecting animate nerve organs networks—were used to predict the AQI dossier on smart places. The Ministry of Environment, Forests, and Climate Change of the Government of India got databases from the CPCB. When it got near guessing Delhi's air quality, the model succeed [12].

### The reasoning of air contamination was projected utilizing the K-Means approach [13]. The equivalence coefficient was computed utilizing physical-opportunity records for contaminants. The KMeans treasure was compared accompanying the Possibilistic Fuzzy C-Means (PFCM) invention. The results accompanied that the upgraded k-resources assembling method created AQI principles accompanying a cut down killing time and better veracity. directed knowledge is the arrangement we use to build guess models. In comparison to other arrangements in their specific classifications, experiments have manifested the superior depiction of decision saplings (categorization), SVR, and stacking collections. For grown regions and places, regression approaches, knowledge, and numerical models were considered.

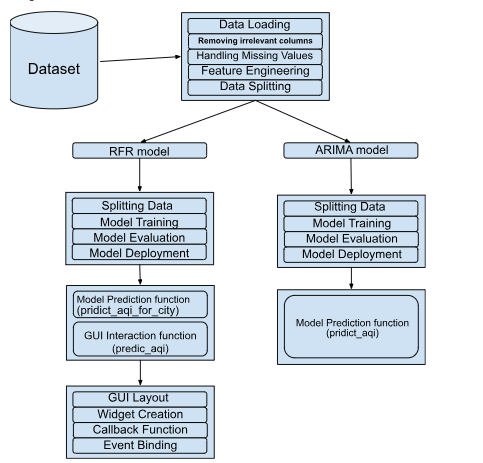
### In this study [14], five recommendation variables—pressure, hotness, initial vapor to water buildup percentage, vapor-particular seriousness, and condensate importance—were used to assemble a novel gadget-education method to estimate the condensate stickiness nearly the wellbore. Nine machine intelligence and composite machine intelligence algorithms were evaluated (GA), containing the novel diversified extreme education structure (MELM), slightest squares support heading machine (LSSVM), and multilayer perceptron, each of that has happened assorted accompanying a atom swarm optimizer (PSO) and genetic invention.

### This study [15] evaluates miscellaneous dossier predicting patterns to foresee the Air Quality Index (AQI) for PM2.5 in Delhi. Employing metrics to a degree R-Squared (R2), Mean Absolute Error (MAE), and Root Mean Square Error (RMSE), the study determines the accomplishment of standalone machine intelligence and deep knowledge models, including LSTM, GRU, LR, KNN, and SVM. Notably, a mixture LSTM-GRU model is projected, professed superior act accompanying an MAE of 36.11 and an R2 advantage of 0.84, indicating allure efficiency in AQI forecast.

### This study fashioned use of PM2.5, individual of the most usual AQI estimate parts. Brown’s burden epidemic affecting average correctly forecasted future Central Jakarta AQI levels established experiment dossier. It prospered better in agreements of precision than the WMA, EMA, and BDES approaches [16].

### They forecasted Dhaka's air status utilizing machine intelligence (ML) models that contained LSTM and added deep knowledge techniques. This approach was new within it working a sole parameter—the everyday hotness—to forecast air pollution [17].

### **Proposed Method:**



**Fig 1: Project Flow**

The primary goal of this research is to formulate a precise and dependable predictive model for projecting the Air Quality Index (AQI) in urban environments. By leveraging machine learning techniques, the aim is to provide timely and precise predictions of air quality levels, which can contribute to better environmental management and public health protection.

1. Approach

Data Gathering: Gathered a dataset [18] from Kaggle that comprises of historical air quality data from diverse urban areas, including pollutant concentrations and corresponding AQI values.

Data Pre-processing: Performed thorough pre-processing on the acquired dataset to handle missing values and inconsistencies. It includes data cleaning, normalization, and feature engineering to make sure the quality and suitability of the data for model training.

Model Selection: Explored various machine learning models suitable for time series forecasting, such as autoregressive integrated moving average (ARIMA) and Random Forest Regression (RFR).

Model Training: Trained the selected model using the pre-processed dataset, apt hyperparameters and model architecture to achieve optimal performance. Utilized techniques such as cross-validation and grid search to fine-tune the model and prevent overfitting.

Model Evaluation: Evaluated the trained model’s performance using unseen test data, assessing its ability to accurately forecast AQI values across different time periods and geographical locations. Measured accuracy using metrics such as mean absolute error (MAE), root mean square error (RMSE), and correlation coefficients.

Model Deployment: Developed a user-friendly interface or application for deploying the trained model, allowing stakeholders such as environmental agencies, policymakers, and the general public to access real-time AQI predictions. Implement monitoring mechanisms to track model performance and incorporate feedback for continuous improvement.

By following this comprehensive approach, the research aims to develop a resilient and scalable predictive model for AQI forecasting, facilitating informed decision-making and proactive measures to mitigate air pollution and safeguard public health.

2. Introduction to Random Forest Regressor (RFR) and Autoregressive Integrated Moving Average (ARIMA):

In machine learning, two powerful techniques frequently employed for forecasting tasks are Random Forest Regressor and Autoregressive Integrated Moving Average (ARIMA) models. Both methodologies offer distinct advantages and can be effectively employed to forecast Air Quality Index (AQI) values.

Random Forest Regressor:

In order to do regression problems, the Random Forest Regressor stands as a versatile ensemble learning method that constructs numerous decision trees during training and generates the average prediction of each tree. It is resistant to overfitting and does a great job managing big datasets with high-dimensional features. The Random Forest Regressor is a useful tool for capturing complex interactions between many contaminants and AQI values in the context of AQI prediction. It is an important tool for AQI forecasting because of its capacity to manage nonlinear interactions and feature importance analysis.

Autoregressive Integrated Moving Average (ARIMA):

ARIMA models are a class of time-series forecasting models that incorporate autoregressive, differencing, and moving average components. These models excel in capturing temporal dependencies and patterns in sequential data, rendering them ideal for forecasting tasks where historical trends are pivotal. In the context of AQI prediction, ARIMA models can adeptly apprehend the temporal dependencies and seasonality inherent in air quality data, such as daily and seasonal variations in pollutant levels. By analyzing historical AQI data, ARIMA models can make accurate predictions about future air quality conditions.

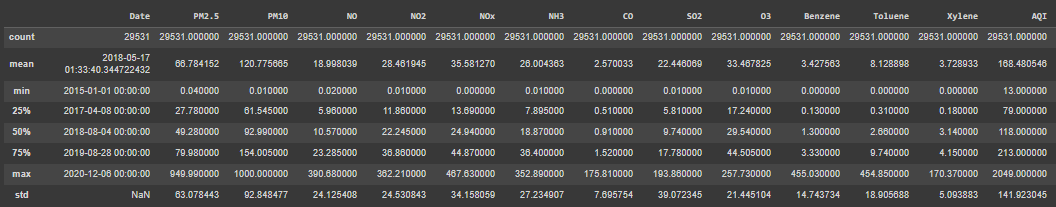
Both Random Forest Regressor and ARIMA models offer valuable insights and predictive capabilities for AQI prediction tasks. By leveraging the strengths of each methodology and integrating them into the research framework, one can develop a robust and accurate forecasting system for predicting AQI values, thus aiding in environmental surveillance and public health administration efforts.

3. Data Preparation and Pre-processing

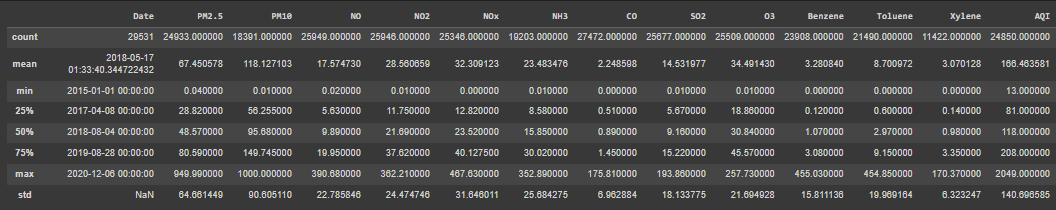
The dataset used here includes missing values, requiring a data cleaning phase. In this phase, the missing values are removed, and the cleaned data is prepared for experimentation. Nevertheless, it's important to highlight that the AQI bucket attribute encompasses six distinct categories. Consequently, following the removal of null values, the distribution of data among these categories may become uneven, particularly for the chosen major cities.

* Pre-processing for Random Forest Regressor:

In our research methodology, rigorous data pre-processing is essential to ensure the integrity and reliability of our analysis. To address missing values in the dataset, we employ a systematic approach that involves imputation. Initially, a duplicate of the original dataset is created to preserve the integrity of the source data. Subsequently, we identify the columns requiring imputation, including key attributes such as temporal information (‘Date’) and various air quality parameters (‘PM2.5’, ‘PM10’, ‘NO’, etc.). A critical step in the pre-processing pipeline is the conversion of the temporal data into a standardized datetime format, enabling consistent temporal analysis across the dataset. Our imputation strategy involves the use of a rolling window approach, where missing values in each attribute are replaced with the mean values calculated over a 5-day interval for each city. This approach leverages both temporal trends and spatial variations in air quality parameters to impute missing values accurately. Finally, any remaining missing values are filled using a backward fill method to ensure the continuity of the dataset. As we can observe by comparative analysis of Fig 2 and Fig 3 we can see that the preprocessing we deployed did not cause a significant change in the mean values over the data set. By meticulously addressing missing data through this comprehensive pre-processing procedure, we aim to enhance the robustness and reliability of our subsequent analysis and modeling efforts.



**Fig 2: Data Analysis before preprocessing**



**Fig 3: Data Analysis after preprocessing**

* Pre-processing for ARIMA:

Apart from the pre-processing applied for the Random Forest Regressor, a few more data processing techniques were applied to make the data appropriate for the ARIMA model. We aggregated the air quality index (AQI) data from the original dataset to facilitate further analysis. Initially, we constructed a pivot table, where the AQI values were organized by date and city. This arrangement allowed us to observe the AQI trends over time for each city individually. Subsequently, we resampled the data to a monthly frequency using the ‘MS’ rule, which enabled us to obtain average AQI values for each city for every month. This resampling process helped in reducing the granularity of the data while preserving the temporal patterns. Furthermore, we computed the average AQI value for the entire country by taking the mean of AQI values across all cities for each month. This aggregated metric, denoted as ‘India\_AQI’, serves as a representative measure of air quality at the national level. By calculating this composite AQI, we gained insight into the overall air quality condition across India over the specified time period. This preprocessing step was essential for synthesizing the individual city-level AQI data into a cohesive and informative dataset for subsequent analysis and modeling.

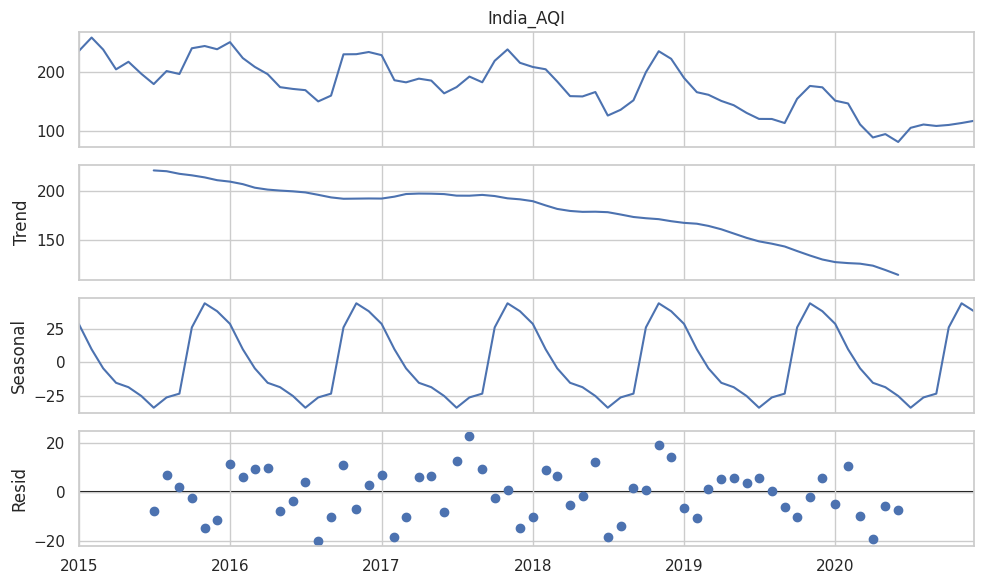
4. Data Splitting and removing stationarity for ARIMA:

We conducted the Augmented Dickey-Fuller (ADF) test to evaluate the stationarity of the time series data depicting the air quality index (AQI) for India. The ADF test serves as a statistical hypothesis examination employed to ascertain the stationary nature of a provided time series. Stationarity represents a fundamental assumption in time series analysis, indicating that the statistical attributes of the data remain consistent over time.

The ADF test on the 'India\_AQI' time series data was executed using the adfuller function from the statsmodels library. The autolag='AIC' parameter signifies that the count of lags employed in the test is automatically chosen according to the Akaike Information Criterion (AIC).

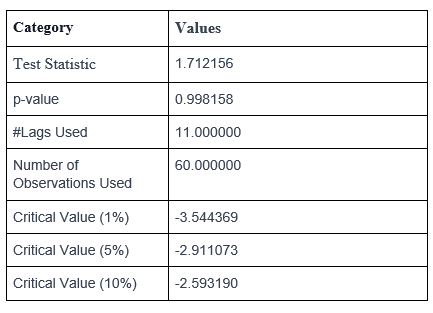
The test results are stored in the dftest variable, which includes the test statistic of the data, number of lags used, p-value of the model, and the number of observations used in the test. Additionally, the dftest attribute contains critical values for different confidence levels.

Subsequently, a pandas Series named df output was created to organize and display the test results in a structured format. The test statistic of the data, number of lags used, p-value of the model, and critical values for significance levels of 1%, 5%, and 10% (Table 1) are included in this Series.



**Fig 4 : Graph for data before differencing**

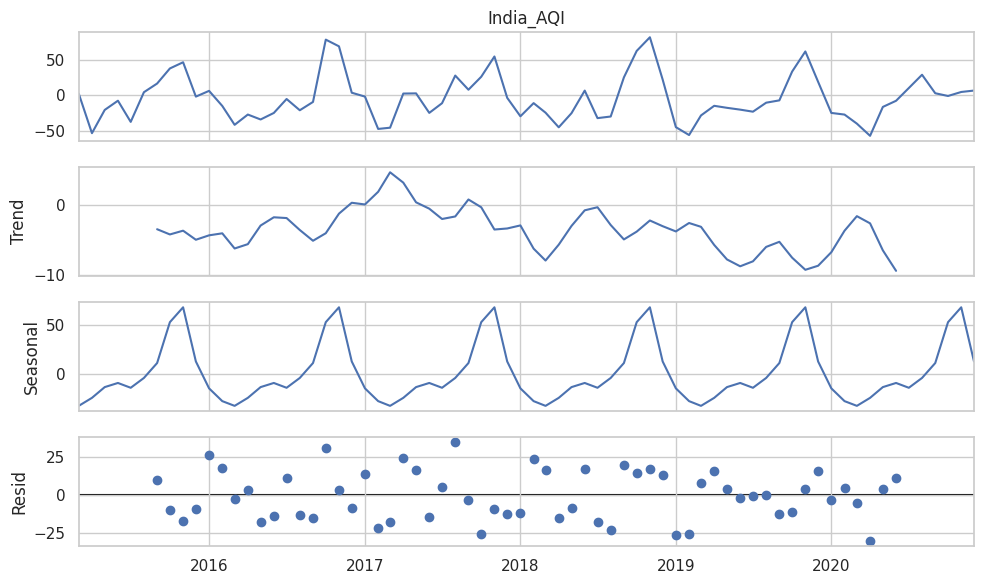
**Table 1:**



Drawing from the acquired p-value (0.998158) (as observed in Table 1), we opt not to dismiss the null hypothesis of the ADF test, suggesting the 'India\_AQI' time series is probably non-stationary (Fig 4 ). This implied that further data transformations or modeling techniques were necessary to achieve stationarity before conducting time series analysis or forecasting.

We then performed differencing on the ‘India\_AQI’ time series data to achieve stationarity. Differencing stands as a prevalent technique in time series analysis aimed at eliminating trends or seasonality from the data, thereby rendering it stationary.

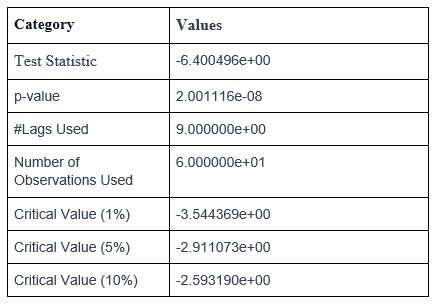
The diff function was employed on the 'India\_AQI' time series data utilizing the parameter periods=2, indicating that we are computing the difference between each observation and the observation two time periods ago. This helps in removing any trend or seasonality that occurs over a two-month period.



**Fig 5 : Graph for data after differencing**

Following differencing, the dropna operation was employed to eliminate any absent values (NaN) that might have arisen due to differencing. This procedure is essential since differencing introduces NaN values for the initial two observations, where differencing computation is not feasible.

**Table 2 :**



On conducting the ADF test again, the p-value came out to be 2.001116e-08 (as seen in Table 2), which implied that stationarity was achieved(Fig 5). Now the data was ready to be split into train and test and to fit the model. It was split according to the dates so as to efficiently split the dataset, train data containing the dataset values from 2015 to 2018 end and the test data containing the rest of the values (i.e. from 2019 to 2020).

5. Model Deployment

* Random Forest Regressor:

Random Forest stands as a predictive modeling technique that amalgamates multiple decision trees. Each tree within the forest is fashioned using independently sampled random vector values, fostering diversity while adhering to the same underlying distribution. Paramount to the Random Forest are its key parameters, namely the number of trees and the number of features. The former dictates the total number of trees within the forest, while the latter specifies the count of features that are randomly selected considered for each decision tree.

The train-test split methodology is employed to partition the data into distinct training and testing sets. In this context, the test\_size parameter, typically set at 0.2, designates the proportion of the dataset allocated for testing, with the remaining 80% earmarked for training the model. Ensuring reproducibility across various code executions is the function of the random\_state parameter. By anchoring the random seed value, this parameter ensures consistent shuffling of data before splitting, thereby yielding reproducible results.

The execution of the RandomForestRegressor involves several vital parameters meticulously defined to adjust the operation of the random forest. 'n\_estimators' controls the quantity of decision trees constituting the forest, established at 200 in this case. 'min\_samples\_split' dictates the smallest number of samples necessary for dividing an internal node during tree formation, with a limit of 2 samples set here. Similarly, 'min\_samples\_leaf' establishes the lowest number of samples required at a leaf node, guaranteeing each leaf includes at least one sample.

The parameter 'max\_features' governs the maximum count of features taken into account for determining the optimal split. Here, the 'sqrt' setting evaluates the square root of the total number of features at each node. 'max\_depth' governs the maximum depth of individual decision trees, constrained to 100 levels to mitigate overfitting. Lastly, the 'bootstrap' parameter, set to False, signifies that the entire dataset is utilized for tree construction without employing bootstrapping.

After parameter setup, the model undergoes training using the fit technique on the provided training data (X\_train and y\_train), encompassing input features (X\_train) and corresponding target variables (y\_train). Upon completion of training, the model is prepared to generate forecasts on new, unseen data.

* ARIMA:

The ARIMA model, is a prevalent option for time series analysis and forecasting. It’s often denoted as ARIMA(p, d, q), where ‘p’ represents the number of autoregressive terms, ‘d’ denotes the degree of differencing needed to make the series stationary, and ‘q’ indicates the number of moving average terms. This model is particularly effective for handling non-stationary data, such as air quality measurements, and is widely used in predictive modeling due to its accuracy.

Within the ARIMA framework, the autoregressive (AR) aspect grasps the correlation between the present observation and its historical data points, whereas the moving average (MA) aspect scrutinizes the model's residual terms. The ‘d’ parameter represents the order of differencing applied to the data to achieve stationarity, which is crucial for accurate modeling.

Typically, second-order differencing is sufficient to make the data smooth and stationary, resulting in the ARIMA(p, 2, q) model. The ARIMA(p, 2, q) model combines autoregressive and moving average components, exhibiting gradual decay characteristics. The values of ‘p’ and ‘q’ are found by using information criteria such as the Akaike Information Criterion (AIC) or the Bayesian Information Criterion (BIC), here we have used AIC.

After getting the data ready for training and testing the model, we took a step-by-step approach to split it into two parts: one for training and the other for testing. Our training set included data from the beginning of our dataset in 2015 up to outcomes at the end of 2018. This gave our model ample time to learn from a wide range of past observations. Meanwhile, the testing set covered the remaining time period, from 2019 to 2020.(Fig ) By dividing the data this way, we could effectively evaluate how accurately our model predicts across different time frames, ensuring that our assessments were both reliable and robust.

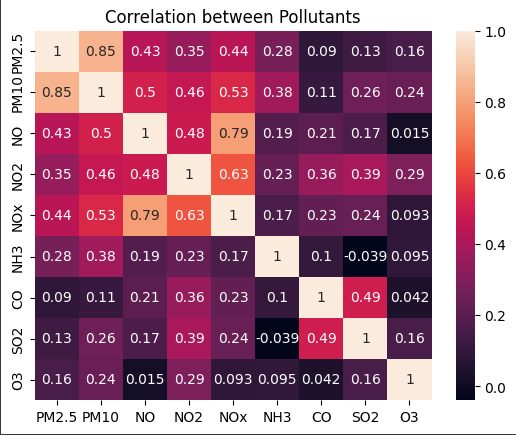
### **Results and Analysis:**

Analysis of the Dataset:

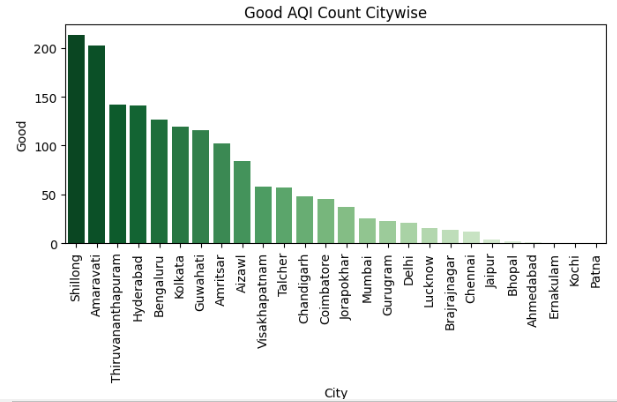
The dataset [18] took advantage of in our model judgment creates from the publicly approachable Air Quality Data in India connecting the age 2015 to 2020, derived from the Kaggle repository. This dataset includes inclusive air feature calculations and Air Quality Index (AQI) readings written on an hourly and regularly support across diversified listening stations located in key cities during the whole of India.

The picked municipalities for reasoning include Ahmedabad, Aizawl, Amaravati, Amritsar, Bengaluru, Bhopal, BrajRajnagar, Chandigarh, Chennai, Coimbatore, Delhi, Ernakulam, Gurugram, Guwahati, Hyderabad, Jaipur, Jorapokhar, Kochi, Kolkata, Lucknow, Mumbai, Patna, Shillong, Talcher, Thiruvananthapuram, and Visakhapatnam.

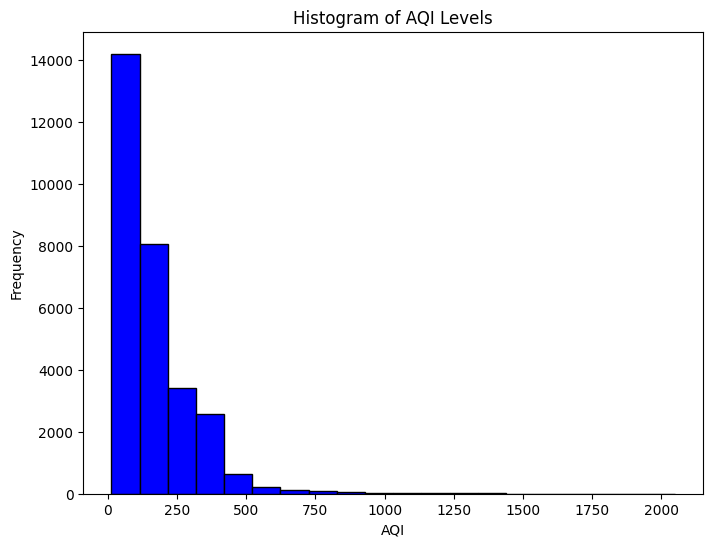
Within the dataset, each city’s records involve differing attributes to a degree date, period, old age, and measured contaminants containing PM2.5, PM10, NO, NO2, NOx, NH3, CO, SO2, O3, Benzene, and Toluene(Fig 6). Additionally, the dataset face AQI readings and categorizes ruling class into six distinct classifications varying from (Fig 7) ‘good’ to ‘harsh’ established predefined thresholds. After thorough study we can visualize that AQI levels pale in 2018-19 but abated in 2020(Fig 9). Which maybe justified apiece evidence that skilled was a India roomy confinement in isolation during that magnitude on account of COVID on account of reduced AQI levels. This classification arrangement acquired immune deficiency syndrome in classification the overall air feature environments for better interpretation and reasoning.



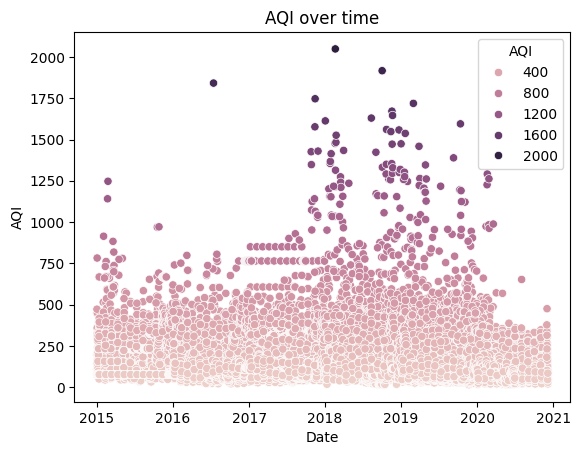
**Fig 6: Correlation between Pollutants**



**Fig 7: No. of times City’s AQI rating was Good**



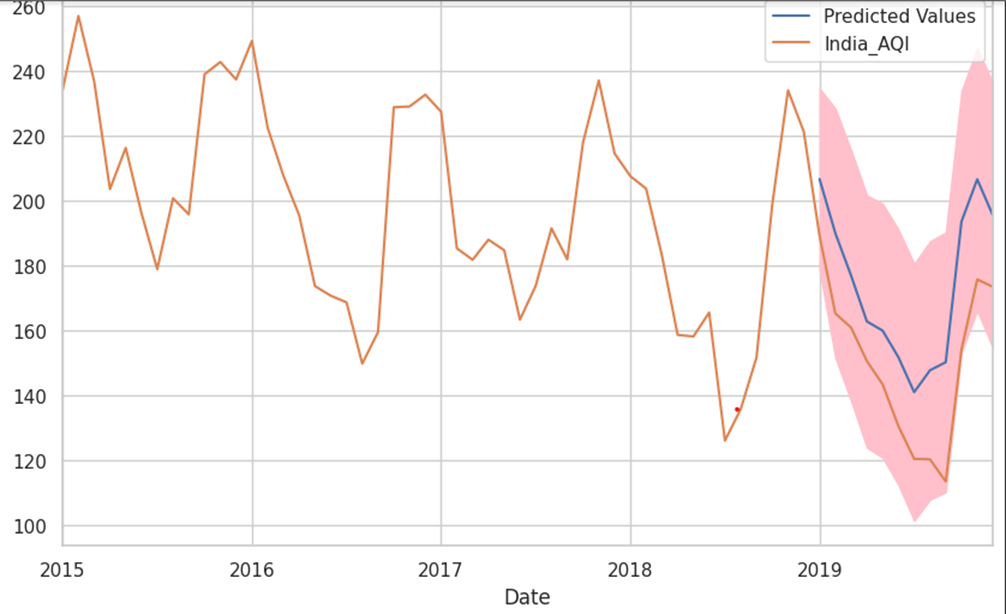
**Fig 8: Histogram showing the Frequency of AQI levels**



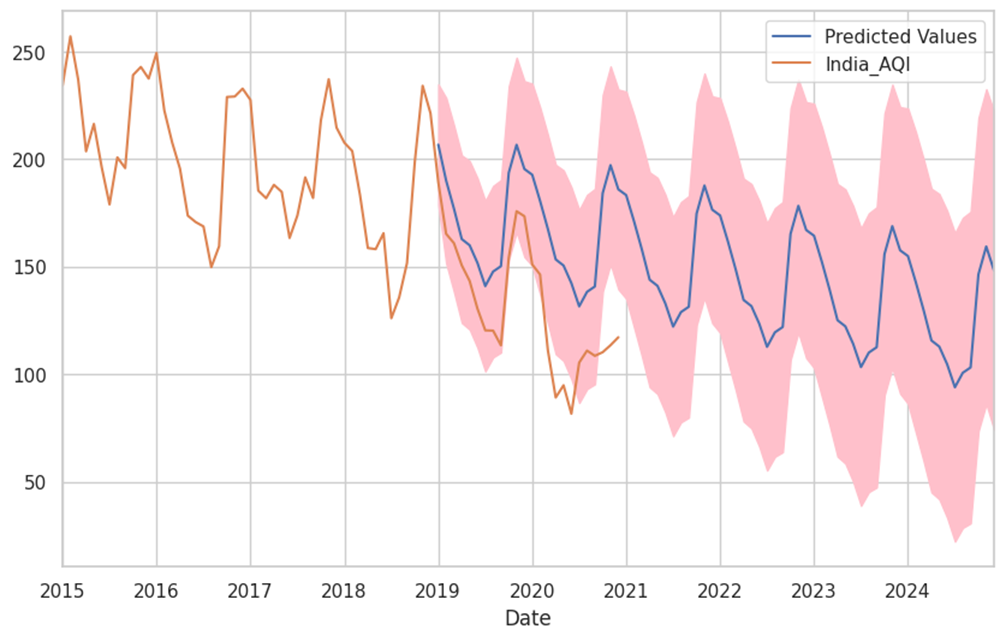
**Fig 9 : Scatterplot of AQI over time**

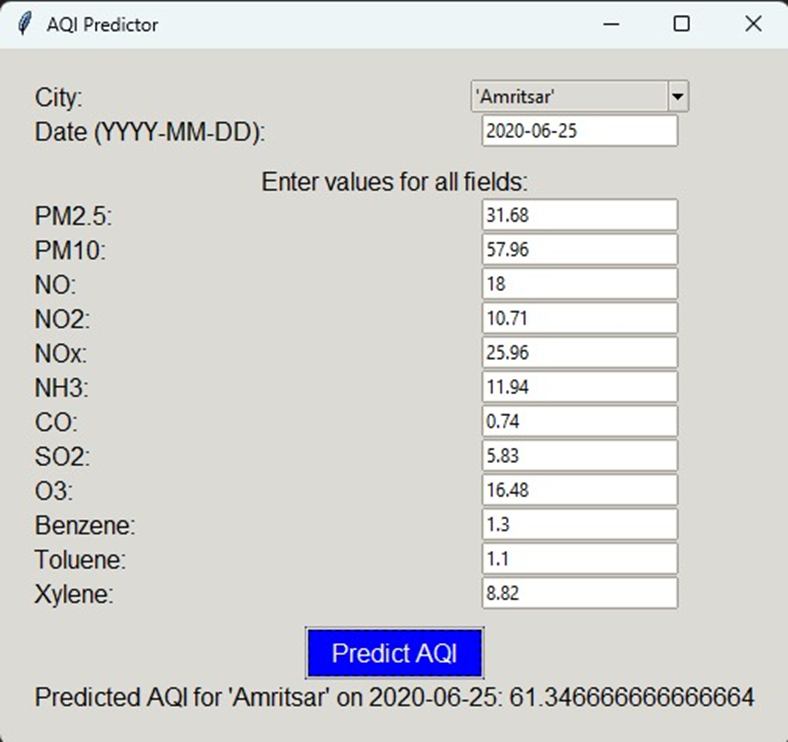
Results from Random Forest Regressor: Our predicting model attained hopeful performance in guessing Air Quality Index (AQI) principles, as proved for one judgment metrics. The Mean Squared Error (MSE) of 1689.24 signifies that, approximately, the agreed distinctness middle from two points predicted and valid AQI principles is comparably depressed, suggesting that our model’s predictions are nearly the real AQI principles. Additionally, the R-regulated (R2) advantage of 0.908 indicates that nearly 90.85% of the difference in AQI principles maybe related by the physiognomy contained in our model. These results display the productiveness of our predictive model in arresting the fundamental patterns and flows in AQI dossier, through facilitating correct predicting of air status levels.

Results from ARIMA Model: The ARIMA model explained promising act in predicting Air Quality Index (AQI) principles, as proved for one evaluation versification. The Root Mean Squared Error (RMSE) of nearly 25.12 wholes signifies that, approximately, the difference middle from two points concluded and real AQI principles is relatively narrow (as visualized in Fig 10 and Fig 11), suggesting that the model’s fore castings approximately join with the real AQI principles. Furthermore, the Mean Absolute Percentage Error (MAPE) of nearly 16.52% displays that, approximately, the model’s predictions avoid the real AQI principles by about 16.52%. These results highlight the influence of the ARIMA model in securing the fundamental patterns and movement in AQI dossier, thereby aiding correct predicting of air characteristic levels.



**Fig 10 : Graph showing predicted Predicted vs Original AQI (ARIMA)**

 **Fig 11 : Graph showing predicted AQI values upto 2024 (ARIMA)**

**Fig 11 : GUI Interface**

### **Future Scope:**

In the realm of air quality prediction and management, the future presents a myriad of opportunities for advancement. One way for improvement is delving into advanced machine learning techniques such as Long Short- Term Memory (LSTM) networks and hybrid models. By harnessing the power of these sophisticated algorithms, we aim to bolster the accuracy and robustness of AQI prediction models, thereby enhancing our ability to forecast air quality dynamics with greater precision. Furthermore, the integration of additional data sources, including weather forecasts, satellite imagery, and demographic data, holds promise for enriching our understanding of the complex factors influencing air pollution. Developing real-time monitoring systems that continuously analyze air quality data and provide immediate alerts to relevant stakeholders, such as government agencies, urban planners, and the general public. Real-time monitoring systems equipped with IoT devices and mobile applications offer a pathway towards continuous air quality analysis. Additionally, geospatial analysis techniques, facilitated by Geographic Information Systems (GIS) tools, can provide valuable insights into spatial variability in air quality, enabling targeted interventions to mitigate pollution hotspots. As we embark on this journey of exploration, we also recognize the importance of conducting comprehensive studies to evaluate the socio-economic and health impacts of air pollution, informing the creation of efficient mitigation strategies and policy suggestions for sustainable urban development. Public awareness campaigns and educational initiatives can also be launched to raise awareness about the importance of air quality and its effects on the human health and the environment. Empowering communities with information and resources can foster collective action towards improving air quality and reducing pollution emissions.

### **Conclusion :**

In conclusion, this research paper has presented a comprehensive investigation into the application of machine learning models, specifically Random Forest Regression (RFR) and Autoregressive Integrated Moving Average (ARIMA), for the prediction of Air Quality Index (AQI). We have utilized these advanced modeling techniques to forecast air pollution levels with considerable accuracy. The results obtained from our experimentation showcase the potential of these models in capturing complex patterns and trends inherent in AQI data, thereby aiding in the formulation of informed policy decisions and mitigation stratMX Clkegies to address air quality concerns. Furthermore, our study highlights the importance of interdisciplinary collaboration between data scientists, environmental researchers, and policymakers to harness the power of machine learning for addressing pressing environmental challenges. As we move forward, further research endeavors should focus on refining existing models, exploring novel algorithmic approaches, and integrating additional data sources to enhance the predictive capabilities of AQI forecasting systems. By embracing these advancements and fostering interdisciplinary cooperation, we can work towards fostering more sustainable environments for the well-being of future generations.

### **References :**

[1] H. Liu, Q. Li, D. Yu, and Y. Gu, “Air quality index and air pollutant concentration prediction based on machine learning algorithms,” Applied Sciences, vol. 9, p. 4069, 2019.

[2] M. Castelli, F. M. Clemente, A. Popovic, S. Silva, and L. Vanneschi, “A machine learning approach to predict air quality in California,” Complexity, vol. 2020, Article ID 8049504, 23 pages, 2020.

[3] G. Mani, J. K. Viswanadhapalli, and A. A. Stonie, “Prediction and forecasting of air quality index in Chennai using regression and ARIMA time series models,” Journal of Engineering Research, vol. 9, 2021.

[4] S. V. Kottur and S. S. Mantha, “An integrated model using Artificial Neural Network (ANN) and Kriging for forecasting air pollutants using meteorological data,” Int. J. Adv. Res. Comput. Commun. Eng, vol. 4, pp. 146–152, 2015.

[5] S. Halsana, “Air quality prediction model using supervised machine learning algorithms,” International Journal of Scientific Research in Computer Science, Engineering and Information Technology, vol. 8, pp. 190–201, 2020.

[6] A. G. Soundari, J. Gnana, and A. C. Akshaya, “Indian air quality prediction and analysis using machine learning,” International Journal of Applied Engineering Research, vol. 14, p. 11, 2019. 24 Journal of Environmental and Public Health.

[7] S. Al-Eidi, F. Amsaad, O. Darwish, Y. Tashtoush, A. Alqahtani and N. Niveshitha, "Comparative Analysis Study for Air Quality Prediction in Smart Cities Using Regression Techniques," in IEEE Access, vol. 11, pp. 115140-115149, 2023, doi: 10.1109/ACCESS.2023.3323447.

[8 Natarajan, S.K., Shanmurthy, P., Arockiam, D. et al. Optimized machine learning model for air quality index prediction in major cities in India. Sci Rep 14, 6795 (2024). https://doi.org/10.1038/s41598-024-54807-1

[9] H. Maleki, A. Sorooshian, G. Goudarzi, Z. Baboli, Y. Tahmasebi Birgani, and M. Rahmati, “Air pollution prediction by using an artificial neural network model,” Clean Technologies and Environmental Policy, vol. 21, no. 6, pp. 1341–1352, 2019.

[10] K. P. Singh, S. Gupta, and P. Rai, “Identifying pollution sources and predicting urban air quality using ensemble learning methods,” Atmospheric Environment, vol. 80, pp. 426–437, 2013.

[11] Y. Zhou, S. De, G. Ewa, C. Perera, and K. Moessner, “Data Driven air quality characterization for urban environments: a case study,” IEEE Access, vol. 6, Article ID 77996, 2018.

[12] U. Mahalingam, K. Elangovan, H. Dobhal, C. Valliappa, S. Shrestha, and G. Kedam, “A machine learning model for air quality prediction for smart cities,” in Proceedings of the 2019 International Conference on Wireless Communications Signal Processing and Networking (WiSPNET), pp. 452–457, Chennai, India, March 2019.

[13] V. Sivakumar, G. R. Kanagachidambaresan, V. Dhilip kumar, M. Arif, C. Jackson, and G. Arulkumaran, “Energy-efficient markov-based lifetime enhancement approach for underwater acoustic sensor network,” Journal of Sensors, vol. 202210 pages, Article ID 3578002, 2022.

[14] A. R. Behesht Abad, S. Mousavi, N. Mohamadian et al., “Hybrid machine learning algorithms to predict condensate viscosity in the near wellbore regions of gas condensate reservoirs,” Journal of Natural Gas Science and Engineering, vol. 95, Article ID 104210, 2021.

[15] Sarkar N, Gupta R, Keserwani PK, Govil MC. Air Quality Index prediction using an effective hybrid deep learning model. Environ Pollut. 2022 Dec 15;315:120404. doi: 10.1016/j.envpol.2022.120404. Epub 2022 Oct 11. PMID: 36240962.

[16] S. Hansun and M. Bonar Kristanda, “AQI measurement and prediction using B-wema method,” International Journal of Engineering Research and Technology, vol. 12, pp. 1621–1625, 2019.

[17] A.-S. Chowdhury, M. S. Uddin, M. R. Tanjim, F. Noor, and R. M. Rahman, “Application of Data Mining Techniques on Air Pollution of Dhaka City,” in Proceedings of the 2020 IEEE 10th International Conference on Intelligent Systems (IS), pp. 562–567, Varna, Bulgaria, August 2020.

[18] Dataset : <https://www.kaggle.com/datasets/rohanrao/air-quality-data-in-india>