Implementing Big Data Analytics for Predicting City-Level Air Quality Index (AQI) and Examining the Impact of COVID-19 using Comprehensive PM and Gas Emission Data

Kalyan Khatry
Herald College Kathmandu
University of Wolverhampton
Kathmandu, Nepal
kalyankhatry2000@gmail.com

Prajwal Adhikari
Herald College Kathmandu
University of Wolverhampton
Kathmandu, Nepal
prajwalad101@gmail.com

***Abstract*—One of the primary issues affecting health and the environment, especially in developing nations like India, is air pollution. Using machine learning techniques, the study analyzes six years of air pollution data from 23 Indian cities, identifying the pollutants that affect air quality and assessing the impact of the COVID-19 pandemic on air quality. This research reveals that short-term lockdowns imposed during the pandemic resulted in a significant drop in pollutants and improved air quality. Support Vector Machine (SVM), Random Forest regression, and XGBoost regression were implemented to predict the Air Quality Index (AQI) values. Random Forest regression outperformed other machine learning algorithms, achieving an accuracy rate of 80%. The AQI values were classified into six categories based on their level of pollution, providing insights into the severity of air pollution in different areas.**

***Index Terms*—COVID-19, Machine Learning, Predict, Random Forest, XGBoost, SVM**

I. Background of the Study

THE presence of harmful substances in the air is referred to as air pollution. The air can be contaminated due to pollutants from various sources, such as natural processes, wildfires, and human activities. However, it is considered that in most urban areas, pollution from vehicles is the primary cause of significant impact on air quality (Kanpur Rani & Vallikanna, 2020). The consequences of COVID-19 have significantly impacted the quality of the air because industry and fuel consumption are the main causes of air pollution. Numerous industries were shut down, and there were much fewer automobiles on the road, during the COVID-19 lockdown. Various air pollutants present in the air contributes to having a negative effect to the air quality. Particulate Matter(PM) is a type of air pollutant created by factors such as burning and dust. NO, NO2 and NOX are all nitrogen oxides and are considered to be a significant contributor to air pollution. Carbon monoxide in produced by uncompleted combustion of fossil fuels which is poisonous to people. Burning fossil fuels like coal and oil releases sulfur dioxide (SO2), which can have negative effects on the respiratory system. Benzene, toluene and xylene are volatile organic compounds that are found in gasoline and solvents. Exposure to high level of these compounds can cause respiratory problems and even cancer (Alattar & Yousif, 2019).

A. Problem Statement

Air pollution has become a global pandemic that is destroying the ecosystem and causing terrible health effects. In the majority of emerging nations, including India, the air quality is getting worse every day. (Jain & Acharya, 2022). Every day, exposure to urban air pollution causes roughly 1,800 deaths in developing cities. Similarly, air pollution is responsible for the premature deaths of 1.2 million people in India. It has been shown that the fuel used for transportation is responsible for 30% of air pollution (Jiyal & Saini, 2020). Agriculture is significantly impacted by air pollution as well. Pollution-related crop production loss has a huge economic impact everywhere in the world. There are projected to be between 11 and 18 billion dollars in annual crop production losses due to air pollution for wheat, corn, and soybeans, with the United States suffering the largest loss ($3.1 billion)(Knowland et al., 2021).

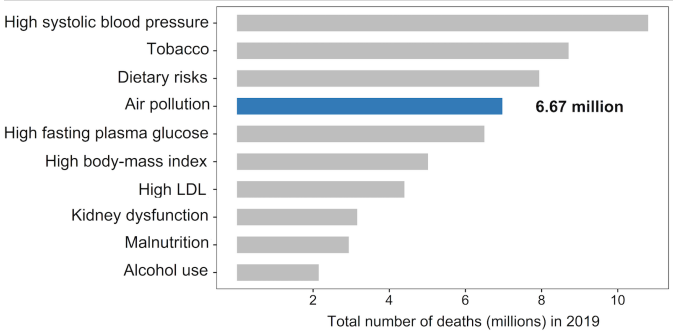

Fig. 1. Impact of Air Pollution, 2019. (Source: State of Global Air:online)

B. Aim/Objective of the work

In this project, we aim to predict the AQI of cities in India using big data analysis techniques. The process will be carried out with the help of particulate matter and gas emission data from urban cities in India. This work's main goal is to effectively create and implement a predictive classification algorithm that can precisely forecast the AQI index based on previous data on emission levels in cities. The project also aims to examine and compare the AQI levels before and after the COVID-19 pandemic. In order to improve public health and air quality, this study aims to give readers a better understanding of the key elements that influence air quality in urban areas.

C. Contributions of the work connected with Methodology

The contributions of the work connected with Methodology are listed as follows,

- Data collection: The dataset is obtained from Central Pollution Control Board (CPCB) which contains records from Jan. 2015 to July 2020. The data contains air quality information about 23 different cities in India.
- Data preprocessing: In order to make sure that the data is appropriate for further analysis, the data from our air quality dataset needs to be preprocessed first.
- Feature Selection: From our dataset, the parameters that affect AQI the most are identified and insignificant columns are dropped. This is done to improve the performance of our model.
- Model selection: Various machine learning models are implemented and then compared to determine the model that produces the best results.

D. Organization of the report

The relevant works on air quality and the effects of COVID-19 are found in the following section. Additionally, it talks about techniques through which AQI can be predicted using machine learning. In the methodology section, the entire work process is described, which includes data cleaning, data processing, and the implementation of different machine learning models. Then, the Results and Discussion section provides a detailed analysis and discussion of this study. The summary of this study is summed up and highlighted in the conculsion section.

II. Related Work

The COVID-19 outburst's effects on air pollution are the subject of recent research. Following the execution of the lockdown strategy to lessen the potential effects of COVID-19, the results indicate a good influence on the air quality. According to the study, compared to the time before the lockdown, PM10 and PM2.5 concentrations were practically halved. The district's air quality in the mechanical and transportation sectors increased by more than 60%. The second and fourth extended periods of lock-down saw a 40–50% improvement in the air quality. NO2, a greenhouse gas emitted during the combustion of fossil fuels, significantly fell as well, with average reductions of 13% overall India and 32.5% in six places(Tyagi et al., 2020).

A recent study looked at the variations in different air quality indicators that were observed during the lockdown period. The authors studied the dataset from District 7 in Florida which covers approximately 8,630 square kilometers. The parameters were observed from 2017 till 2020 which allowed the authors to determine how the lockdown period affected the variables. Due to data limitations, the researchers additionally performed Pearson correlation statistical tests to determine the relationship between the parameters for specific countries. According to the study's findings, O3 levels declined at a rate of about 7-8% on average across the nine air quality stations that were looked at, and the patterns in the air pollution indicators were shifting. (El Traboulsi et al., 2022).

This paper analyzes air quality data from Delhi region and predicts AQI level of the region through the use of various supervised machine learning algorithms. The dataset contains 45,000 records of all states of India but the authors only analyzed the values of the Delhi region which had around 5000 records. In order to predict the AQI level, algorithms such as Linear Regression, Lasso Regression, Random Forest Algorithm, and Decision Tree were used and the accuracy of 97%, 97%, 95% and 100% were achieved respectively. The authors suggest that when a large dataset is used in the prediction process, random forest algorithm should be used (Pal et al., 2021).

The authors used the air quality data of Henan Province taking the average daily amount of six main pollutants as characteristics factors, AQI as a decision factor, and combined multiple regression algorithms in machine learning, established multiple regression models that can forecast the AQI. The RFR algorithm and GBR algorithm were more suitable for air quality prediction, and these two algorithms have the advantages of high parallelization, strong noise resistance and high accuracy. The authors also suggest that the model could be improved by simply obtaining more data and correcting the prediction output of inconsistent distribution or other methods (Li et al., 2021).

For this study, concentrations of various pollutants and harmful gasses for various Indian cities were analyzed. For the comparison of air quality across various cities, the National Air Quality Index (NAQI) was used. It was found that vehicles and automobiles were the major common cause of generation of pollutants in monitored cities. Due to high energy consumption during winter seasons, the concentration of pollutants was also found to be higher during that time. The study concluded that the levels of SPM and RSPM were extremely high in all cities of India (Nasir et al., 2016).

The authors of this study compare and analyze how four of India's most industrialized and polluted states are affected by important air quality parameters. Data for the study was extracted from the CPCB website utilizing an automatic monitoring network for various time periods. The study focused on the AQI monitoring locations that had roughly 60% of the valid data for the time periods used. One of the drawbacks of

this study is that it only looks at trends in metropolitan areas, not in regional or rural areas (Nandhini et al., 2021).

This study forecasts the air quality in Indian cities using a variety of machine learning methods. The dataset is prepared for maximum accuracy using machine learning techniques like LR, SVM, NB, K-NN, RF, and DT which provides an accuracy of 98%, 95%, 99%, 70%, 97% and 100% respectively. The decision tree technique is effective at reducing outliers and unnecessary data, which makes it a helpful tool for forecasting air quality issues. The process of this air quality forecasting system can be automated by showing the prediction result in a website or desktop application, and it can be updated for prospective usage in an artificial intelligence setting (Reddy et al., 2021) .

III. Methodology

A. Data Collection

India is considered to be one of the most polluted and populated countries in the world. And some of the cities in India are the most polluted cities in the world. These cities have also become a threat in terms of air pollution. The poor air quality of air could lead to several health and environmental issues. Some of the main causes of air pollution include automobiles, factories, waste burning, and other numerous variables. The Central Pollution Control Board (CPCB) of India keeps track of the information crucial to the deterioration of air quality. And, in this research, the dataset was collected from CPCB. The dataset contains data from Jan. 2015 to July 2020, with 16 features and 29,531 rows. This dataset has data from 23 different cities. Table 2 provides a brief description of the pollutant's particles and AQI from the dataset.

	PM2.5	PM10	NO	NO2	NOx	NH3	CO	SO2	O3	Benzene
Count	24933	18391	25949	25946	25346	19203	27472	25677	25509	23908
Mean	67.45	118.127	17.574	28.56	32.309	23.483	2.248	14.531	34.491	3.28
Std	64.661	90.605	22.785	24.474	31.646	25.684	6.962	18.133	21.694	15.811
Min	0.04	0.01	0.02	0.01	0	0.01	0	0.01	0.01	0
25%	28.82	56.255	5.63	11.75	12.82	8.58	0.51	5.67	18.86	0.12
50%	48.57	95.68	9.89	21.69	23.52	15.85	0.89	9.16	30.84	1.07
75%	80.59	149.745	19.95	37.62	40.1275	30.02	1.45	15.22	45.57	3.08
Max	949.99	1000	390.68	362.21	467.63	352.89	175.81	193.86	257.73	455.03

	Toluene	Xylene	AQI
Count	21490	11422	24850
Mean	8.7	3.07	166.463
Std	19.969	6.323	140.696
Min	0	0	13
25%	0.6	0.14	81
50%	2.97	0.98	118
75%	9.15	3.35	208
Max	454.85	170.37	2049

Fig. 2. Parameters Table

B. Pre-Processing

The quality of the data that we have is the most important factor for proper analysis, visualization, and creation of efficient machine learning models. The block diagram of our proposed model is given in Figure 3.

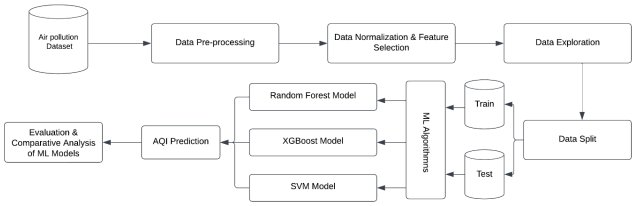

Fig. 3. Block diagram of proposed model

Pre-processing of the data helps in speeding up and generalizing the performance of machine learning algorithm models. Having empty/missing values and outliers are some of the major problems in extracting the features of the data. The preprocessing technique uses several operations on the data. Filling out the NAN values, removing the outliers, changing the outliers, applying the logarithm to each value in the columns, or calculating the interquartile range in each feature can be performed while preprocessing the data. Observing the dataset, Xylene has the highest number of missing values with 61.32% (18109 out of 29531) whereas CO has the least number of missing values with 6.97% (2059 out of 29531). The skewness of the dataset was measured to measure the asymmetry in a dataset. It was found that the variables PM2.5, NO, NH3, CO, SO2, Benzene, Toluene, Xylene, and AQI are highly positively skewed, and PM10, NO2, and O3 are also positively skewed but to a lesser extent. Therefore, the missing values in the dataset are filled with the median values.

In order to increase the performance of the ML models, normalization of data is also carried out. For example, the dataset is obtained from different stations, and the representation of dates in the dataset is different. Therefore, the data is scaled up into a symmetrical representation throughout the experiment using the Python built-in library.

C. Feature Selection

The dataset includes the AQI parameter, which measures the quality of air. Additionally, the National Ambient Air Quality Standards list a total of six categories for determining the AQI. (i.e., the AQI bucket). The air quality index can be divided into 6 different categories (Kumar et al., 2021). The table below shows the categorical value of AQI.

AQI Bucket	AQI Value
Good	0-50
Satisfactory	51-100
Moderate	101-200
Poor	201-300
Very Poor	301-400
Severe	401-500

Fig. 4. AQI Table matrix

Correlation between each of the features is determined between the input and the target variable. In this model, we

have AQI as our target variable and other pollutants factor as our input variable. The correlation between each of the features can be seen in Fig 4. The AQI and other features are correlated, and the pollutants for which the correlation value exceeded the cutoff point of 0.4, i.e., the correlation was strongly positive, are detected. The variables with the maximum correlation between the target feature are then further analyzed to fit into the model. It can be observed that PM10, PM2.5, CO, NO2, SO2, NOX, and NO have higher correlations and are major factors affecting the value of AQI. Although O3 does not have higher correlation with AQI but it is one of the major pollutant factor causing air pollution so this feature is not dropped and the remaining feature columns are dropped for further analysis.

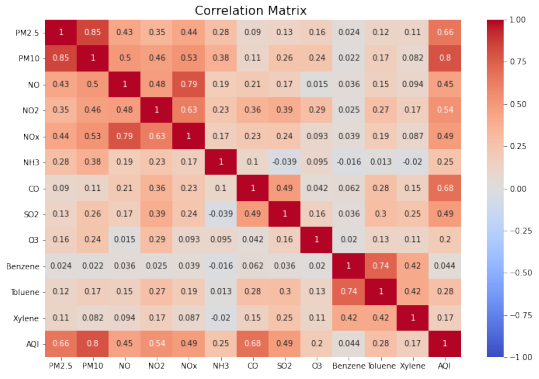

Fig. 5. Correlation matrix

Machine learning algorithms do not perform well while having outliers. Therefore, the outliers in the dataset are observed. Then the data are first introduced to logarithmic transformation to reduce the impact of outliers by normalizing the differences. And finally, Inter Quartile Range (IQR) is used to remove the outliers present in the data making it ready to fit in the model. Doing so will improve the performance of the models used. The distribution of data after removing the outliers is given in the figure below.

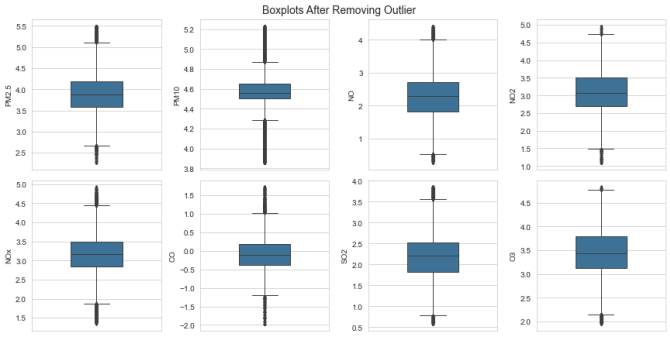

Fig. 6. Box plots after removing outlier

D. ML Algorithms

The data is splitted into test data and train data. Train data are the data that are trained or fed to the model for predicting the AQI. The test data is the unseen data. How well the model works on the unknown data is typically determined by this dataset. In our model, we have splitted data in 80:20 ratio. 80% data are splitted into training data and the remaining data are the test data. Performance of three different algorithms are compared on the same dataset.

1) Random Forest: Random Forest makes predictions by using numerous decision trees. As a result of its excellent accuracy and capacity for handling huge datasets, it is a popular machine learning method. The algorithm constructs several decision trees based on random subsets of the input features. When all of the decision trees projections are combined, it provides the final prediction.

2) SVM: A popular supervised learning technique for regression problems is SVM. In order to divide the data into two classes, the best hyperplane must be found. The algorithm looks for a line that best fits the data. High-dimensional data and both linear and nonlinear data can be handled by SVM (Kulkarni et al., 2022).

3) XGBoost: XGBoost is an algorithm that generates predictions using decision trees. Although it is similar to Random Forest, it boosts the performance of each decision tree individually using a gradient method. The method gradually adds decision trees to the model, each tree seeking to correct the problems of the one before it. Competitions and real-world applications frequently make use of XGBoost because of its potential for handling complicated datasets (Nababan et al., 2022).

IV. Result and Discussion

This section focuses on the experimental design and empirical analysis of predicting AQI (Air Quality Index) values based on air pollutants. To evaluate machine learning models, the air pollution dataset is split into training and testing subsets. Various data processing tasks are carried out using Python libraries such as Pyspark, Scikit-learn, Seaborn, etc.

A. Data Exploration

This section of the study provides insight into patterns present in the dataset. Exploratory data analysis is an important initial step in data analytics that is typically carried out before any machine learning models are applied. This step involves analyzing various important aspects of the data to better understand its characteristics. During this process of analysis for air quality index data, we explore the following key factors:

- The trends and the patterns in the pollutants level over six years (2015 - 2020)
- The level of pollutants in the air and identifying the most polluted cities and their average Air Quality Index (AQI) values.
- Estimating the most significant pollutants that contribute to increased AQI values.

By performing these analyses, we gain important insights into the characteristics of the air quality data, which can help inform subsequent steps of the data analysis process, including selecting appropriate machine learning models and pre-processing steps. Fig 7 provides insights into the linearity between the input features and the target feature.

In the figure below, we can see that the features selected for predicting the value of AQI are linearly dependent. Being linearly dependent is also a major reason for choosing these features. The dataset is explored to determine the overall AQI value with respect to pollutants that significantly contribute to its increase. Fig 8 displays a timeline graph of pollutants directly responsible for higher AQI values. As evident from the graph, the levels of each pollutant increase and decrease annually, with values varying across different years. It can be noticed that PM2.5 and PM10 show a slight decrease after. After 2018, the SO2 level started to increase, whereas the O3 level remained somewhat constant from 2018 to 2020. Similar trends are observed in Benzene, Toluene and Xylene levels. However, CO is the only pollutant that does not exhibit seasonal variations.

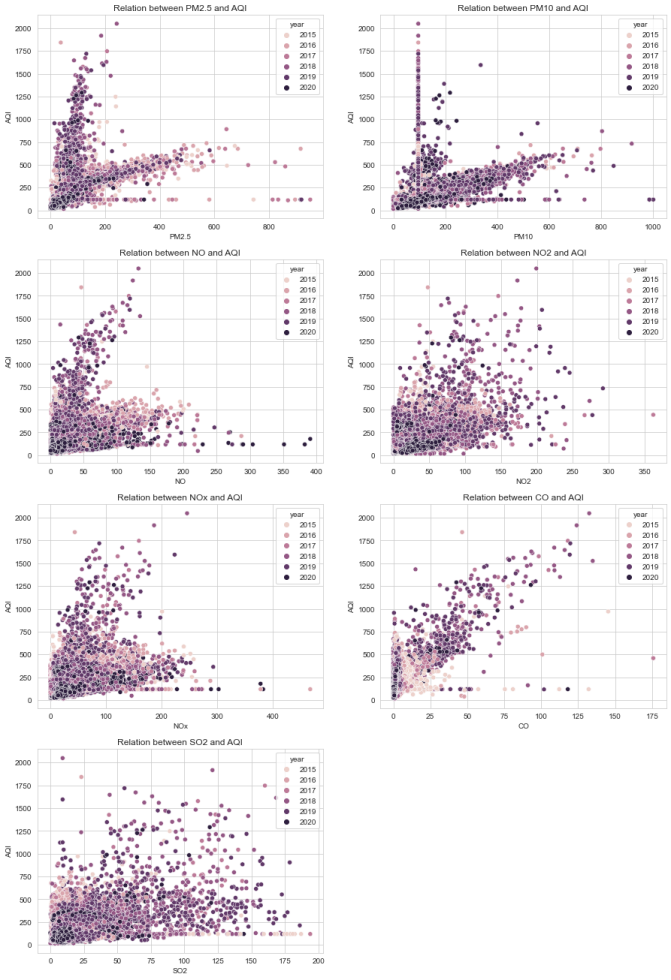

Fig. 7. Relationship between different gasses and AQI

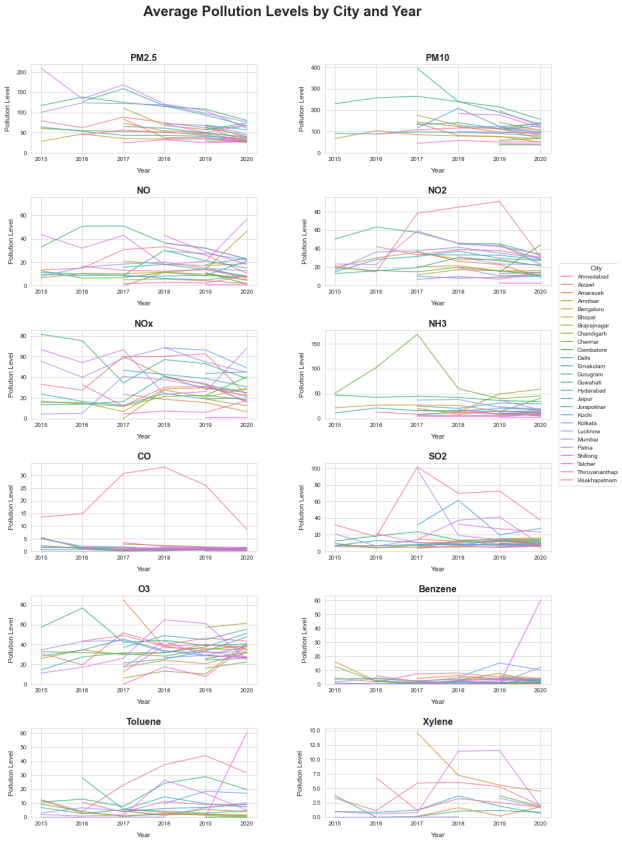

Fig. 8. Average Pollution Levels by City and Year

B. *Impact of COVID-19 on AQI*

This study indicates that one of the main environmental issues we are now facing is air pollution. Particulate matter (PM), ozone (O3), carbon monoxide (CO), sulfur dioxide (SO2), nitrogen oxides (NOx) and other pollutants are among the many that we breathe in. These pollutants are released from a variety of sources including fossil fuel , industrial activities, transportation, and many more. And inhaling these pollutants will have a negative impact on our health, including respiratory and cardiovascular conditions, as well as the environment.

In this dataset, we analyzed the AQI values from 2015 to 2020 and observed that the AQI values had a similar pattern till 2019. However, during the lockdown period in 2020, we observed a significant decrease in the AQI values. This observation indicates that the lockdown had a positive impact on air quality and reduced the level of pollutants in the air.

In India, a number of things contribute to air pollution, such as vehicle emissions, industrial pollutants, building activity, and crop residue burning. One of the primary causes of air pollution in many Indian cities is vehicle emissions. The level of pollutants like PM and NOx has increased as a result of the continued development in the number of vehicles on the road. Industrial emissions are also a significant contributor to air pollution, with industries releasing pollutants such as SO2, NOx, and PM. The burning of crop residue is a another signifi-

cant element in India's air pollution. Farmers burn crop residue after harvest, which releases pollutants such as PM and CO into the air. This practice is common in India and contributes significantly to air pollution. And during the lockdown period, the vehicles were not allowed to run, and the industries were all closed. Also, travelling to any destinations were band and the peoples were forced to sit in their household. They were not able to harvest crops or burn any kind of crop residue. Therefore, these could be the major reason for the decrease in the AQI of these Inidan cities during the lockdown period.

The AQI values observed in this dataset indicates that the lockdown period had a positive impact on air quality and reduced the level of pollutants in the air. However, it is important to note that there are various factors contributing to air pollution in India, and a concerted effort is required to address this issue. The government, industries, and citizens must work together to reduce emissions and promote sustainable practices to ensure clean air for future generations.

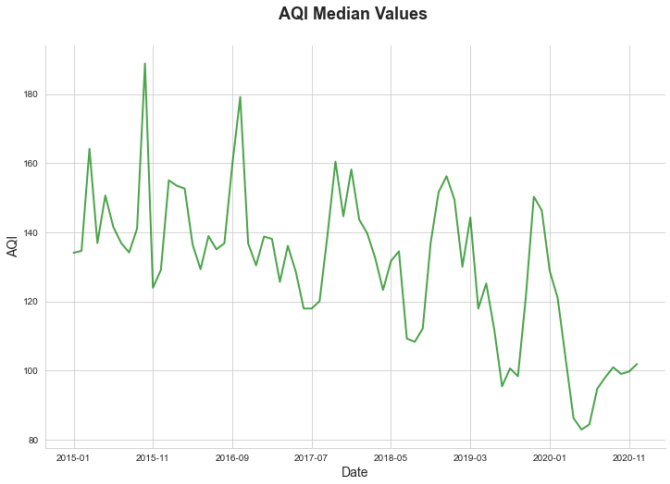

Fig. 9. AQI Median Values

C. Discussion

A particular matter influences the air pollution. And it is very important to know the factors that are highly influencing the quality of air. And knowing the factors that are highly influential factors of air pollution should be identified as they can be valuable information to improve the air quality. In this research, Feature importance analysis was performed using machine learning to find out which pollutants had a major impact in predicting the AQI levels. Our results indicate that PM2.5 is the most important feature for the model in predicting the AQI, followed by PM10, CO, O3, SO2, NO2, NOX, and NO.

Fig 10 shows the importance of each feature with PM2.5 having the highest importance score of 0.5. PM10, CO, and O3 also had relatively high importance scores, with values of 0.2, 0.1, and 0.07, respectively. SO2, NO2, NOX, and NO had lower importance scores, with values ranging from 0.01 to 0.03.

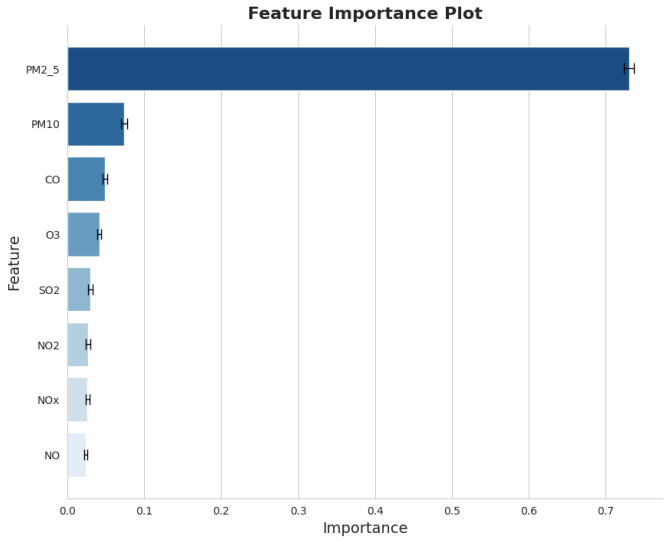

Fig. 10. Feature Importance Plot

Our findings are consistent with past studies that have demonstrated that PM2.5 both significantly affects human health and contributes to poor air quality. It is worth noting that our analysis was based on a specific dataset and machine learning model. According to the data set and model that are being utilized, the relative value of each attribute may change. Nonetheless, our results provide valuable insights into the factors that influence AQI levels and can be used to guide future research and policy decisions. Our research shows that PM2.5 is the most important feature for predicting AQI levels, followed by PM10, CO, and O3. These results can be used to develop effective strategies to improve air quality and protect public health.

D. Best Fit Model

Air Quality Index (AQI) takes into account several factors such as ozone, particulate matter, and carbon monoxide levels. Accurately predicting the AQI is crucial for managing air quality and protecting public health. In this study, three different ML algorithms were trained to predict the value of AQI: Random Forest, XGBoost and SVM.

In this study, Random Forest outperformed SVM and XGBoost in predicting the AQI. Random Forest achieved an accuracy of 80%, while SVM achieved 71% accuracy and XGBoost achieved 77% accuracy. This could be due to Random Forest's ability to handle large datasets and its ability to capture complex interactions between the input features. SVM may have underperformed due to its reliance on finding a linear hyperplane and its difficulty in handling large datasets. XGBoost may have slightly underperformed due to the possibility of overfitting the data, especially if the number of decision trees is high.

The results of this research demonstrated that Random Forest was the most reliable model for AQI prediction. It is essential to understand that the models' effectiveness can

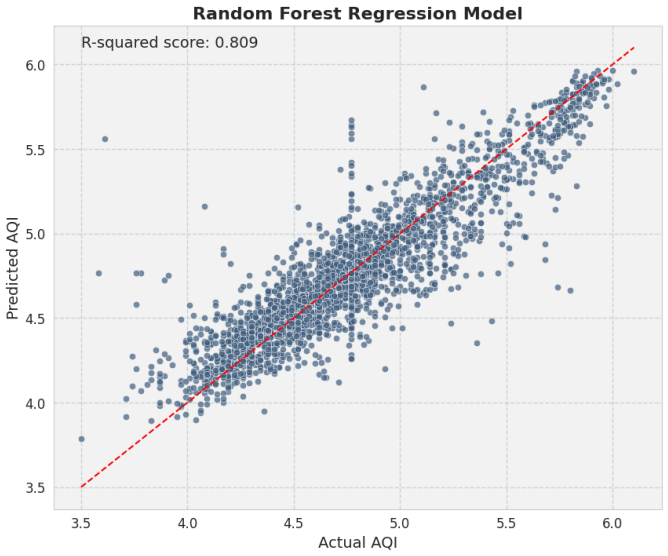

Fig. 11. Random Forest Regression Model

change based on the particular dataset and issue at hand. So, all models are compared before selecting the best one for a given problem. The actual value and the predicted value using Random Forest model is shown in the figure below.

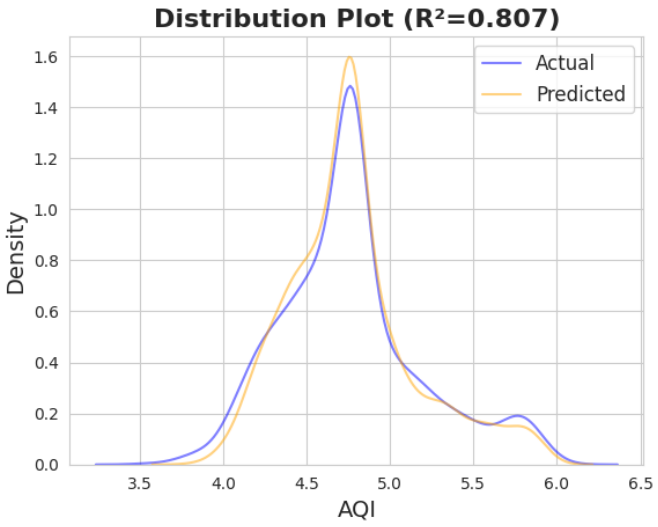

Fig. 12. Distribution Plot

Mean square error (MSE) and R-Square score are the two most common metrics used to evaluate how effective a regression model is. The average squared difference between the predicted outcomes and the actual values is what the MSE calculates. A lower MSE value implies that the model has a better match to the data. Looking at the experiment performed, the MSE value of our model was 0.032 which suggests that the model has relatively low error rate and is a good sign.

The R-Square Score is a statistical measure of the proportion of variance in the dependent variable that can be anticipated from the independent variables. The model is better at explaining data variability the higher the R-Square value. The R-squared score of 0.809 tell us that 82.8% of the variability in the data can explained by the model. This suggests that the model is a good fit for the data and can make accurate predictions.

The MSE and R-squared score offers important insight into the regression model's performance and may be used to predict a model's accuracy and make the necessary adjustments to enhance its performance.

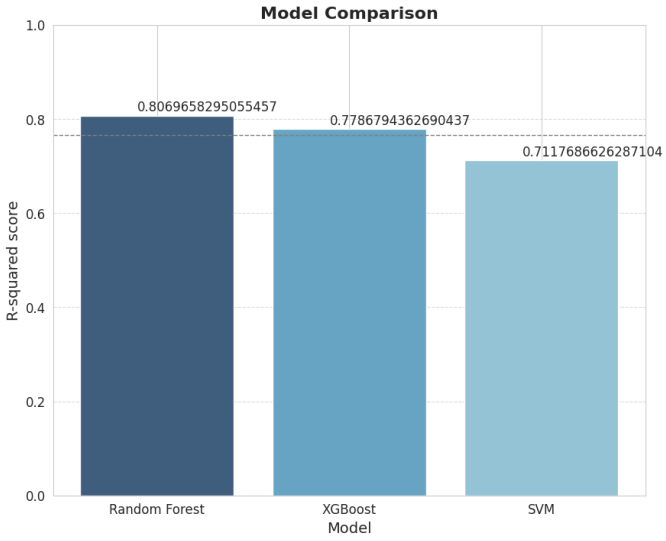

Fig. 13. Model Comparison

V. Conclusion

Air pollution has a major negative influence on human health, making it a global problem. Air quality index (AQI) values, which are used to gauge the quantity of air pollution, are predicted in this study using machine learning algorithms. Three different algorithms - Random Forest, SVM, and XGBoost - are compared to predict the AQI values. Based on the degree of pollution, the predicted AQI values are divided into six categories: Good, Moderate, Satisfactory, Poor, Very Poor, and Severe. The effect of COVID-19 and the ensuing lockdown on air quality has been investigated in this study. And the findings demonstrated that the reduction in human activity including transportation and industrial operations during the lockdown led to a considerable increase in air quality. Further analysis was also performed to determine the important factors affecting air quality. The results showed that PM2.5 is the most important factor of the dataset followed by PM10, CO, O3, SO2, NO2, NOX and NO. The machine learning algorithms used in this research have achieved varying levels of accuracy. The Random Forest algorithm has proved to be the most effective, achieving an accuracy rate of 80.6%, while SVM has achieved 71.1 % accuracy and XGBoost has achieved

77.8% accuracy. These algorithms may help in monitoring and regulating the level of air pollution as well as increasing the precision of AQI predictions.

VI. Acknowledgement

We would like to thank Mr. Basudeo Shrestha and Mr. Jnaneshwar Bohara for their valuable feedback, guidance and support throughout the entire process. We would also like to express our gratitude to the University of Wolverhampton for providing us access to academic papers and resources necessary to conduct this study.

References

[1] Rani, V.K. and Vallikanna, A.L., 2020, November. Air Pollution Monitoring System using Internet of Vehicles and Pollution Sensors. In 2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA) (pp. 249-252). IEEE.

[2] Jain, E. and Acharya, D., 2022, July. Mobile Sensing and Modeling Air Pollution Hotspots in Urban Neighborhoods. In 2022 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT) (pp. 1-6). IEEE.

[3] Jiyal, S. and Saini, R.K., 2020, November. Prediction and monitoring of air pollution using Internet of Things (IoT). In 2020 Sixth International Conference on Parallel, Distributed and Grid Computing (PDGC) (pp. 57-60). IEEE.

[4] Knowland, K.E., Keller, C., Ott, L., Pawson, S., Saunders, E., Wales, P., Duncan, B., Follette-Cook, M., Liu, J. and Nicely, J., 2019. Near real-time air quality forecasts using the NASA GEOS model (No. GSFC-E-DAA-TN70165).

[5] Tyagi, A., Kharb, L. and Chahal, D., 2020, December. Scrutinizing Patterns of Air Pollution in India. In 2020 2nd International Conference On Advances In Computing, Communication Control And Networking (ICACCCN) (pp. 5-9). IEEE.

[6] El Traboulsi, Y., Khalifa, I., Abuzwidah, M., Shanableh, A., Al-Ruzouq, R. and Hamad, K., 2022, February. Effect of COVID-19 Lockdown on Traffic-related Air Pollution in Florida, USA. In 2022 Advances in Science and Engineering Technology International Conferences (ASET) (pp. 1-6). IEEE.

[7] Pal, S., Pramanik, D. and Jain, E., 2021, November. Effectiveness of Machine Learning Algorithms in Forecasting AQI. In 2021 International Conference on Technological Advancements and Innovations (ICTAI) (pp. 492-495). IEEE.

[8] Li, C., Li, Y. and Bao, Y., 2021, November. Research on air quality prediction based on machine learning. In 2021 2nd International Conference on Intelligent Computing and Human-Computer Interaction (ICHCI) (pp. 77-81). IEEE.

[9] Nasir, H., Goyal, K. and Prabhakar, D., 2016. Review of air quality monitoring: case study of India. Indian Journal of Science and Technology, 9(44), pp.1-7.

[10] Nandhini, C., Mirthul, E.S. and Dhurandher, B.K., 2022, July. An assessment on Air Quality in various Polluted Industrialized states of India due to COVID-19. In 2022 1st International Conference on Sustainable Technology for Power and Energy Systems (STPES) (pp. 1-6). IEEE.

[11] Reddy, Nagarjuna, and P. Selvi Rajendran. "The Prediction of Quality of the Air Using Supervised Learning." In 2021 6th International Conference on Communication and Electronics Systems (ICCES), pp. 1-5. IEEE, 2021.

[12] Alattar, N. and Yousif, J., 2019, December. Evaluating Particulate Matter (PM2. 5 and PM10) Impact on Human Health in Oman Based on a Hybrid Artificial Neural Network and Mathematical Models. In 2019 International Conference on Control, Artificial Intelligence, Robotics & Optimization (ICCAIRO) (pp. 129-135). IEEE.

[13] Kumar, R.S., Arulanandham, A. and Arumugam, S., 2021, October. Air quality index analysis of Bengaluru city air pollutants using Expectation Maximization clustering. In 2021 International Conference on Advancements in Electrical, Electronics, Communication, Computing and Automation (ICAECA) (pp. 1-4). IEEE.

[14] Nababan, A.A., Zarlis, M. and Nababan, E.B., 2022, October. Air Quality Prediction Based on Air Pollution Emissions in the City Environment Using XGBoost with SMOTE. In 2022 IEEE International Conference of Computer Science and Information Technology (ICOSNIKOM) (pp. 1-6). IEEE.

[15] Kulkarni, M., Raut, A., Chavan, S., Rajule, N. and Pawar, S., 2022, August. Air Quality Monitoring and Prediction using SVM. In 2022 6th International Conference On Computing, Communication, Control And Automation (ICCUBEA (pp. 1-4). IEEE.

[16] State of Global Air (2019) Impacts on Your Health — State of Global Air. Available at: https://www.stateofglobalair.org/health (Accessed: 05 May 2023).

Kalyan Khatry was born in Bhojpur, Nepal and is a student of computer science at Herald College Kathmandu. He was a awarded AAA scholarship for his excellent academic performance and is also the StAR of his batch. His research interests includes artificial intelligence, deep learning and data analysis.

Prajwal Adhikari was born in Kathmandu, Nepal and is currently studying computer science at Herald College Kathmandu. He was a awarded AAA scholarship for his excellent academic performance. His research interests in big data and data analysis.

VII. Appendix

All the code used in this study are hosted on GitHub and can be accessed through this link