Insights from Data Science and Machine Learning: Understanding Global Air Quality Dynamics

Sai Shradha V [1,†], Kukatlapalli Pradeep Kumar[2,†] and P Vijaya [3,†]

[1,2] *Department of Computer Science and Engineering, Christ University, Bangalore, India.*
[3] *Department of Mathematics and Computer Science, Modern College of Business and Science, Muscat, PC 133, Sultanate of Oman.*

Abstract

Air pollution is one of the growing global phenomena which is affecting all living organ-isms on the planet and is one of the major climatic crisis that has to be addressed efficiently in a sustainable manner. Air pollution is mainly rising due to the increasing amount of ozone, carbon monoxide and Particulate Matter 2.5 because of various reasons including emission of these gases from industries, burning of fossil fuels, increasing number of aero-sols in the air which has ultimately led to global warming and ozone layer depletion. This research paper mainly aims to focuses on analysis and interpretation of the impact of increasing levels of Ozone, Carbon monoxide, Particulate Matter 2.5 and other pollutants which contributes to global air pollution trends across various major cities around the world including Indian cities. Various hypothesis tests had been performed such as One-way Analysis of Variance tests, to demonstrate the relationship between the impact of contaminants, primarily carbon monoxide, ozone, Particulate Matter 2.5 on the overall Air Quality index various major cities around the world. The effect of the major pollutant on the overall Air Quality Index of the most of the countries has been identified and various machine learning algorithms like K-Nearest Neighbor, Logistic regression, LSTM, Random forests have been implemented to compare and analyze the impact of the air pollutants. This re-search also contributes to United Nations Sustainable development goals on Climate action to study and reduce the increasing global warming which leads to increase in temperature, melting of ice caps day by day.

Keywords

Data Analytics, Global Warming, Air Quality Index, Carbon monoxide, Statistical tests

1. Introduction

Human wellbeing is greatly impacted by the indoor conditions, since ninety per-cent of people spend more than ninety percent of their time indoors. Every year, indoor air pollution, also known as indoor air pollution, causes 3.8 million deaths. It can be caused by the activities of occupants, including cooking, smoking, using electronics, and emitting materials from buildings. Due to shifting life-styles and urbanization, research on air quality control has moved from outdoor to indoor settings. Human health may be adversely affected by decreased IAQ, which may result in illnesses linked to buildings. Approaches for regulating and improving IAQ, as well as monitoring systems, are crucial. This paper offers a thorough analysis of the main sources of INDOOR air pollution, control tactics, health implications, health problems, and trends in surveillance and control Finding the sources of air pollutants is essential to managing indoor air quality effectively.

The overall condition of the outside air, the activities of people inside buildings, and the furnishings, appliances, and building materials are among the primary variables that affect indoor air quality Given that pollutants can travel from the outside into the interior, it is well known that the degree of air sealing in constructions and the quantity of pollutants outdoors have a significant impact on the condition of indoor air (indoor air quality). The contaminants outside include Particulate matter, also called particulates (Particulate Matter), dioxide of carbon (CO2), Sulphur oxides (SO2),

AISD-2024: Second International Workshop on Artificial Intelligence: Empowering Sustainable Development, October 2, 2024, co-located with the Second International Conference on Artificial Intelligence: Towards Sustainable Intelligence (AI4S-2024), Virtual Event, Lucknow, India.

sai.shradha@btech.christuniversity.in (Sai Shradha V); kukatlapalli.kumar@christuniversity.in (Kukatlapalli Pradeep Kumar); pvvijaya@gmail.com (P Vijaya)

0009-0000-0661-8088 (Sai Shradha V); 0000-0002-8893-4312 (Kukatlapalli Pradeep Kumar); 0000-0002-2117-9543 (P Vijaya)

oxides of carbon (CO), oxides of nitrogen (NO2), and the products of combustion are among the emissions that are present in in-door air setting. Most importantly, ventilation system design and operation have a significant effect over the indoor air quality. Air circulation helps keep up good indoor air quality and foster healthy conditions by exchanging stagnant indoor air with fresh outdoor air. There are several advantages to running a building's ventilation system [1], such as: (a) supplying clean air and oxygen that people require to breathe; (ii) bringing indoor air pollution levels down to levels that fall within short-term exposure limits for dangerous contaminants; and (iii) eliminating vapors and Odors. Exhaustive Literature review is provided with concepts to biological pollutants, Sulphur and Nitrogen elements. This is followed by Results and Discussions which includes Data Visualization, Statistical tests and machine learning observations.

1.1. Objectives

The main objective of this research is to identify the various air pollutants in the atmosphere and how to contributes to the decrease in Air Quality Index all over the world. Various tests and analysis have mainly been done as to observe the relationship as well as the influence of these air pollutants on the Air Quality Index of the atmosphere. Hence various machine learning algorithms have been chosen specially to identify the impact of Ozone on the atmosphere, which indirectly contributes to overall Air Quality Index value of a place. After various statistical analysis it has been found out the significance of ozone on the Air Quality , hence various algorithms have been extensively performed to under-stand its influence. The accuracy level of the algorithms is also clearly mentioned , if which the highest is the random forest with which we can directly identify he relation and the dependence of the major pollutant Ozone on atmospheric as well as climatic conditions.

2. Literature Review

Yang and Zhao's review article from 2023 offers a thorough examination of the air quality models (AQMs) that are used to simulate ozone (O_3) levels in China, emphasizing the models' capabilities and uses. After reviewing 212 pa-pers released since 2010, the authors found that, whereas CMAQ, WRF-Chem, and NAQPMS exhibit less definite biases, models like GEOS-Chem and CAMx frequently overestimate O3 concentrations. The main causes of long-term O_3 rise, according to the study, are anthropogenic emissions, especially those from industry and traffic. Short-term pollution episodes are increasingly influenced by meteorological factors. To increase simulation accuracy, the scientists stress the need for better emission inventories and model methods. Additionally, they talk about the consequences of the present NO_x-focused control measures, which can unintentionally raise O_3 levels in areas with low VOC concentrations. The subsequent review primarily addresses the different types of air pollutants, such as particulate matter and aerosols like nitrogen oxides, Sulphur dioxide, carbon oxides, and ozone which contributes to the global air pollution.

2.1. Particulate matter

A carbon-based particle connected to reactive metals and organic com-pounds is called a particulate matter . Due to its potential for inhalation, which can harm the heart and lungs, it poses a serious health risk. Indoor activities and outdoor particles are two reasons why indoor levels of frequently surpass outdoor ones. A few of the factors that cause indoor are cooking, burning fossil fuels, smoking, using machines, and engaging in recreational activities at home. Because Particulate Matter, a particulate matter derived from the burning of fossil fuels, can enter small airways and alveoli, it presents a greater risk to health. The primary sources of interior particulate matter come from cooking and cigars smoking; cleaning activities contribute significantly less.

One quarter of indoor concentrations are caused by other human activities. In every city but Bengaluru, Particulate Matter 2.5 and Aerosol optical depth show a decreasing trend from January to May. Although there was more daily variability and less steep trends, satellite-based and Aerosol optical depth revealed trends, satellite-based and Aerosol optical depth revealed trends that were comparable to surface measurements. Bengaluru saw a significant de-crease in Particulate Matter 2.5 concentrations during the lockdown, with daily readings as low as thirteen and a half microgram with m−3. and Aerosol optical depth decreased to almost half of its previous year's level. Delhi saw high levels of particulate matter in the month of January, but in 2020, concentrations of Particulate Matter 2.5 were roughly fourteen percent lower than they were prior to the lock down [1].

2.2. Nitrogen Oxides

Two of the main nitrogen oxides linked to sources of combustion like heaters and cooking stoves are nitrogen oxides (NO) and nitrogen dioxide (NO2). The local sources and sinks of nitrogen have an impact on the ambient concentrations of NO and NO2. Under ambient conditions, NO2 is regarded as a primary pollutant; however, NO2 quickly oxidizes to form NO2. Nitrous acid (HNO2) is a powerful oxidant that is produced when NO2 com-bines with water. It is a common indoor pollution. Both indoor and outdoor sources can raise or lower indoor NO2 levels, and the separation between buildings and roads has a major impact. [2]

Cities like Bangalore, Delhi, the city of Kolkata, and Pune, Maharashtra have high seasonality in their nitrogen dioxide (NO2) concentrations, which de-cline from January to May. The main causes of this seasonality are variations in mixing layer depth and lifetime. All six cities saw a notable decrease in NO2 concentrations after the imposition of stringent security precautions in March 2020, with Delhi exhibiting the greatest decrease. These declines are similar to those observed in Chinese and European cities and correlate with the falls in tropospheric nitrogen dioxide segments detected by Ozone monitoring instrument [3].

2.3. Sulphur Dioxide

The most prevalent and important gas among the Sulphur oxides in the atmosphere is Sulphur dioxide (SO2), which is mostly produced by utilizing fossil fuels. It mixes with Particulate Matters and aerosols to create a complex mixture of unique air. Another major source of indoor SO2 is outside air. Oil furnaces, aired gas-powered devices, smoke from cigarettes, kerosene heaters, coal or wood stoves, and other indoor sources are among those that release SO2. Because SO2 is reactive, indoor surfaces can easily absorb it, and indoor levels of the gas are typically lower than outdoor levels. Buildings usually have hourly SO2 concentrations of less than 20 parts per billion, and human exposure to SO2 predominantly impacts the way we breathe by means of inhalation. Industries, which includes thermal electric power plants, is responsible for more than eighty per-cent of SO2 emissions; through the lockdown, average Sulphur dioxide amounts in Chennai and Delhi dropped somewhat. Pune saw a fairly stable two-year peri-od whereas Bengaluru and the city of Kolkata experienced moderate increases of thirty-one percent and forty-two percent, respectively. During the lockdown period of time, Mumbai had a significant spike in mean SO2 concentrations (81percentage) with levels ranging from around 21 microgram m−3 to a high value of 91 [4].

2.4. Carbon Monoxide gas

Carbon monoxide (also known as CO)is an odorless, colorless gas that is mainly out over processes of combustion like heating or cooking. Furthermore, it could enter indoor spaces through outside air. Gas space heaters, unvented kerosene heaters, leaky burners and furnaces, gas water heaters, wood cookers, fireplaces, stoves with gas, generators and cigarette smoke are the primary sources of indoor atmospheric carbon monoxide (CO) emissions. The typical Carbon monoxide. Concentrations vary from half to five parts per million in a structure without gas stoves. Thus, when the building is close to a gas stove, the level of carbon dioxide can be as high as 30 parts per million. Carbon monoxide (CO) being exposed can have detrimental effects on one's health, including effects on the nervous system and the heart. It can also increase the risk of death [5].

2.5. Ozone

Cities' concentrations of ozone (O3) varied daily and did not exhibit a steady trend over the course of the analysis. Except for Kolkata and Pune, concentrations were, nevertheless, somewhat lower than the mean of the preceding year (2017– 19). While Pune possessed higher concentrations up until the third week of March 2020, throughout 2020, Kolkata recorded ozone concentrations that were higher than the baseline. It is reasonable to anticipate a minor rise in volatile organic compound (VOC) emissions, particularly in circumstances akin to Covid-19 where stringent community disinfection protocols are implemented to curb the disease's transmission. Nonetheless, NOx emissions in Kolkata decreased, which is in line with a forty seven percent de-crease in NO2, but it's possible that a chemical regime limited by VOCs enhanced the formation of O3. When exposed to three thousand parts per mil-lion of carbon dioxide (CO2), headaches become more intense, drowsiness, fatigue, and concentration problems occur.

2.6. Biological pollutants

Biological allergens including pollen, cockroaches, house dust, and animal dan-der are examples of biological pollutants that can be found in indoor environments. Moreover, bacteria, fungus, and viruses are examples of microorganisms that contribute to biological pollutants indoors. These allergens come from fungi, animals, or insects and can trigger allergic reactions when certain antibodies bind to them. Furry pets, home dust insects and mites' mound, vegetation, pests and rats are a few examples of indoor allergen sources in addition to external ones. Sensitivity, infections of the respiratory tract, breathing allergic diseases, and wheezing fits can all result from viral and bacterial exposure. Exposure to bacteria and viruses indoors can lead to a variety of adverse health is-sues , which includes infectious and not transmissible ones. Following the guide-lines for managing asthma can help to lessen the effects, but taking global action is necessary to reduce exposure and improve asthma outcomes. It is necessary to conduct more research to look into dual or various exposures and to find respiratory disease patterns that could increase an individual's vulnerability to atmospheric pollution [7-9].

3. Problem Statement

There are innumerous number of pollutants present in the atmosphere which has severe effect on the atmosphere. Hence it is very necessary to identify and analyze the pollutants present in the atmosphere. The major problem is to identify the pollutants that have the most impact on the Air Quality Index of different regions.

A wide range of pollutants, such as Particulate Matter , carbon dioxide , sulfur oxides , carbon monoxide , nitrogen oxides , and ground-level ozone , are contributing to a becoming complex and interconnected web of environmental degradation. Particulate Matter, using its fine particles, not only impairs air quality but additionally leads to soil and water contamination, influencing biodiversity and ecosystem functions. Air pollution has emerged as one of the most pressing global environmental challenges with profound implications for ecosystems, climate, and human health. Rising carbon dioxide levels are hastening climate change, causing extreme weather, disturbances to natural habitats, and changes in temperature patterns. Acid rain's primary ingredients, sulfur and nitrogen oxides, are destroying built environments, aquatic systems, and soils. This has a domino impact on cultural heritage and food security. Because ground-level ozone is created by photochemical interactions between NOx and volatile organic com-pounds (VOCs), it can harm flora, lower crop yields, and have an adverse effect on the health of forests. Meanwhile, by encouraging the continued presence of other greenhouse gases, carbon monoxide's impact on the atmospheric composition exacerbates climate change. When combined, these pollutants are causing a number of environmental problems, such as ocean acidification, desertification, loss of biodiversity, and diminished natural resilience. Fighting global air pollution necessitates a coordinated and diversified response because of the interconnection of these pollutants and their diverse sources, which range from farming and urban development to industrial emissions and transportation. The overall environmental effects of these pollutants will be examined in this study, with particular attention paid to their interconnections, and accumulated effects on climate and ecosystems, mainly Air Quality Index. In doing so, it hopes to further knowledge of the worldwide air pollution situation and provide guidance for policies aimed at reducing its far-reaching effects.

4. Methodology

Here we conceptualize various graphs including bar graphs, histogram, scatter plots, statistical tests like One-Way ANOVA tests, and machine learning model implementations. We have collected Global air pollution dataset from an open source which consists of 23464 rows of various cities air quality data with variables like Country, City, Air Quality Index Value, Air Quality Index Category, Carbon Monoxide Air Quality Index Value, Carbon Monoxide Air Quality Index Category, Ozone air quality index Value, Ozone Air Quality Index Category, NO2 Air Quality Index Value, NO2 Air Quality Index Category, Particulate Matter 2.5 Air Quality Index Value, Particulate Matter 2.5 air quality index Category.

4.1. Dataset Quality, Completeness and potential biases

This study's dataset, which focuses on global air pollution, includes detailed geo-located data for major pollutants like carbon monoxide , nitrogen dioxide , ozone , and particulate matter , along with corresponding Air Quality Index values and classifications for cities worldwide. The dataset is useful for examining pollutant distributions and trends in air quality, but in order to make sure that the conclusions are reliable and correct, it is crucial to talk about its completeness, quality, and any potential biases.

The fact that the dataset was sourced from an open-source platform has ad-vantages and disadvantages. Open-source datasets are helpful for extensive analysis because they are generally widely available and span a large geographic area. However, the manner in which the data were gathered, handled, and pre-served may have an impact on the data quality. The overall trustworthiness of the dataset may be impacted by, among other things, disparities in monitoring equipment, inconsistent data recording techniques, or different data collection intervals between regions [10].

The dataset may have biases due to a number of causes. For example, the majority of monitoring stations for air quality are situated in or close to metropolitan areas, which causes an overrepresentation of areas with greater amounts of pollutants. The results could be skewed overall if rural areas or locations with fewer monitoring resources are underrepresented. A number of pre-processing procedures, including data cleaning, were carried out to address problems with data quality, accuracy, and biases. Outliers and erroneous values were found and either fixed or eliminated in order to increase the confidence of the results. In order to ensure more dependable results and reduce noise in the dataset, this step was crucial.

4.2. Dataset Quality, Completeness and potential biases

In order to validate the statistical analysis, thorough explanations of the ANOVA and the cross-tabulation findings must be given. An ANOVA may reveal significant differences in AQI when the AQI is categorized by specific pollutant levels, for example, while a cross-tabulation may reveal correlations between categorical variables like AQI categories and pollutant thresholds. It will support the discussion of these findings in the context of machine learning outputs, highlighting the complementary nature of both approaches in highlighting significant factors influencing air quality changes. By highlighting these statistical results, the paper's conclusions about the effects of various pollutants, particularly ozone, on AQI are strengthened [11].

The following section gives an insight into and visualization of the impact, frequency and effect of various pollutants on the environment through Pie charts, bar graphs, histograms, scatter plots and null hypothesis tests like Pearson correlation, One-way ANOVA test and Chi-squared tests.

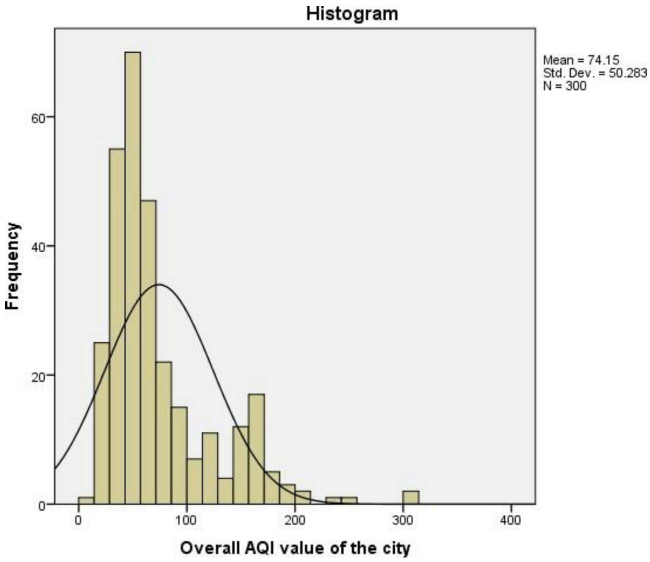

Figure 1: Histogram of overall air quality index of the cities

The above Fig. 1. Shows the frequency distribution of histograms of Air Quality Index of various metropolitan cities of different countries throughout the world acquired from the dataset

Variable taken refers to the different categories of the air quality based on air quality index values. The variable taken here refers to the air quality index value of Carbon Monoxide and their frequency.

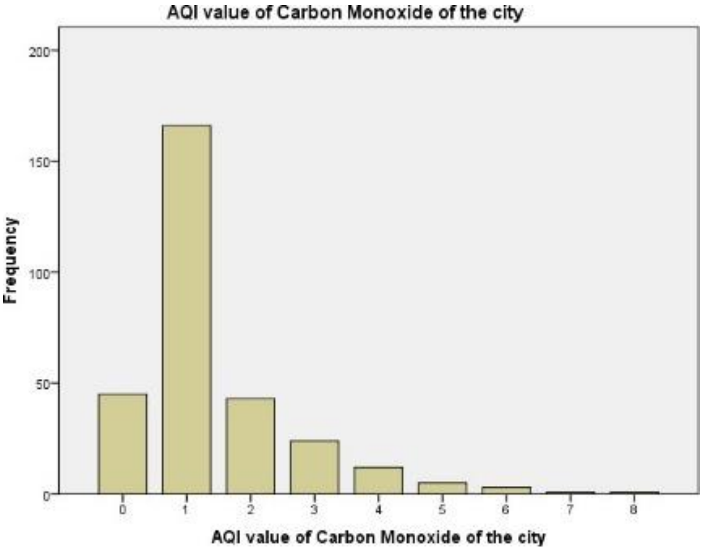

Figure 2: Bar graph of air quality index value of carbon monoxide of the city

The above Fig. 3. Shows bar graph depicting the frequency of Air Quality index value of carbon monoxide over various cities present as major air pollutant in the atmosphere.

A. Scatter Plot

X-axis Variable is General the air quality index reading for the city.
Y-axis Variable is the city's carbon monoxide Air Quality Index value

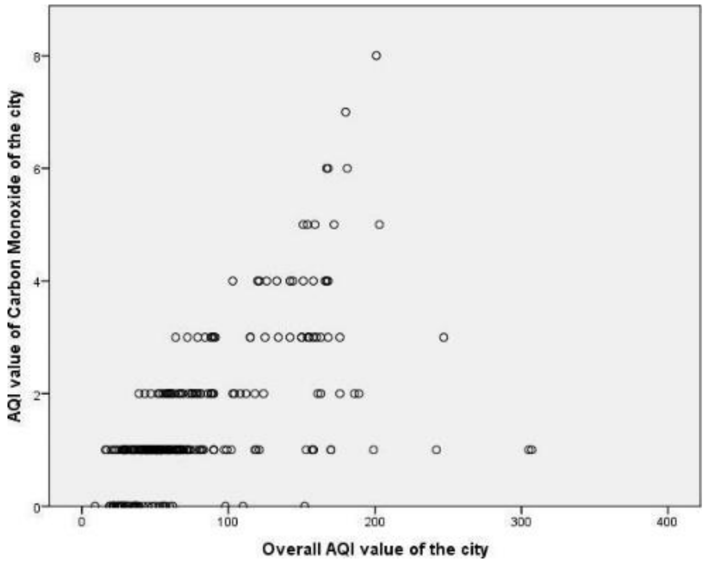

Figure 3: A scatter diagram showing city’s overall the air quality index value and its carbon monoxide the air quality index value

The above Fig. 4. Shows a scatter plot on the ratio of Carbon Monoxide to the Overall Air Quality Index of major cities throughout the world. Histogram of overall air quality index of the following section depicts various statistical tests including various null hypothesis tests performed on respective parameters which gives the relation between various variables and values associated with them [10].Here the skewness and kurtosis analysis are done on the Overall Air Quality Index value of the city .Since, the value of kurtosis is, 2.935+0.281=3.216, where it is greater than 3, it is a leptokurtic curve , when the respective graph is plotted city.

Table 1
Correlation Among the Total Air Quality Index and Air Quality Index Value Of Carbon Monoxide

	Pearson Correlation	Asymp. Sig. (2-sided)
Overall air quality index value of cities	0.984295061448353	.144
Air Quality Index Value of Ozone of The City	300	.103

The above Table 1. Shows the Pearson correlation between two variables such as Overall Air Quality Index value of the city and Air Quality index value of Ozone of the cities which has a positive correlation of 0.984295061448353The null hypothesis tests are performed on the on the city's overall air quality index and the carbon monoxide index: Firstly, Chi-Square Tests ,and One-Way ANOVA tests.

Table 2
Chi-Square Tests on the overall air quality index value and Carbon Monoxide air quality index value

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	3.878	2	.144
Likelihood Ratio	4.537	2	.103

The above Table 2. depicts the Chi-Square tests on the two variables, the Air quality index category of Carbon Monoxide and the Air quality index cate-gory of Nitrogen Dioxide.

The p-value is 0.144 which is greater than 0.05 shows that there is no significant association between air quality index category of Carbon Monoxide of the city and air quality index category of Nitrogen Dioxide of the city. One-way Anova Dependent list: Overall Air Quality Index value of the city Factor: air quality index category of the city (Categorical Variable).

Table 3
One-way ANOVA on the overall air quality index value and Carbon Monoxide air quality index value

	Total Squares	df	Sig. (2-sided)
Between Groups	565.729	1	.011
Total	749.429	6	

The above Table 3. shows the One-way ANOVA Test between Overall Air Quality Index value of the city and Air Quality Index Category of the city which is a categorical variable having categories good, bad, hazardous etc. Dependent list taken here is Overall Air Quality Index value of the city and the Factor is Air Quality Index category of the city.(Categorical Variable).

5. Results

5.1. Machine Learning model perceptive

The capacity of the machine learning models—Logistic Regression, Random Forest (RF), LSTM, KNN, and KNN—to identify intricate patterns and correlations in the data on air pollution led to their selection. Random Forest is an efficient method for determining the most significant contaminants influencing AQI since it can handle nonlinear interactions well and provides insights into feature relevance. The ability of LSTM networks to describe temporal connections is particularly helpful because air quality data is time-series in nature. Because of its ease of use and ability to capture local trends, K-Nearest Neighbors (KNN) was added [12]. These characteristics can be important when analyzing the effects of particular contaminants in different locations. Even though it's straightforward, the benchmark model for binary classification applications, such as classifying AQI levels, is logistic regression. The amalgamation of these models facilitates an all-encompassing examination of temporal patterns and pollutant interactions, augmenting the comprehensive comprehension of ozone's impact on AQI. By performing various correlation analysis, it has been found that there is a very positive correlation between Ozone air quality index Values , Particulate matter and the Overall air quality index value of the city which is 0.984295061448353

Various Machine Learning Algorithms are implemented such as K-Nearest Neighbor, Logistic Regression , Long short-term memory and Random Forests.

Random Forests:
Training variables is ozone air quality index Values and Particulate matter

Testing variables is Overall air quality index value
Mean Absolute Error obtained is 0.01 degrees and Accuracy level 99.98 percent-age

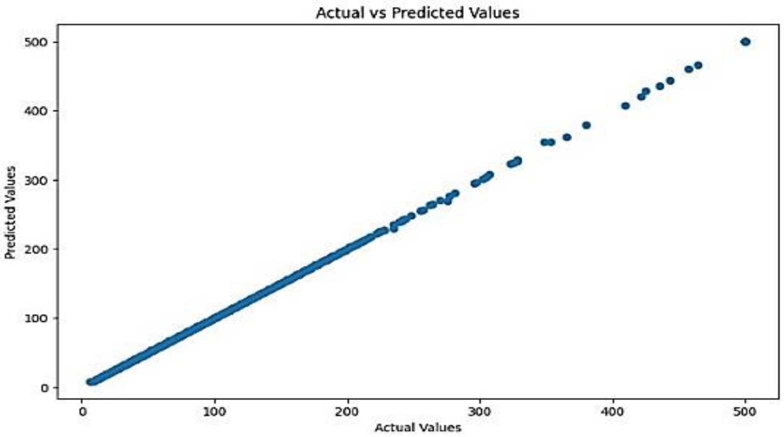

Figure 4: The exponentially increasing graph between the actual versus predicted values.

The above Figure 4. Shows the exponential growth in the accuracy of the implementation of the Random Forests algorithms based on the features of Predicted versus Actual values.

Table 4
Result and Accuracy scores of the dataset implemented using KNN

Heading level	Values			
Baseline K-Nearest Neighbors	0.83	4693		
accuracy	0.85	0.46	0.44	4693
macro avg	0.45	0.46	0.44	4693
weighted avg	0.86	0.85	0.85	4693
Mean Square Error for Nearest Neighbor	13.726			
Root Mean Square Error	3.705			
Baseline K-Nearest Neighbors	0.785			
Scores(Cross validate) for K-Nearest Neighbors model	[0.73086144 0.72638645 0.72234655]			

The above Table 4. Shows brief results of the implementation of K-Nearest Neighbor algorithm on Ozone air quality index Values, Particulate matter and Overall air quality index Value of the cities.

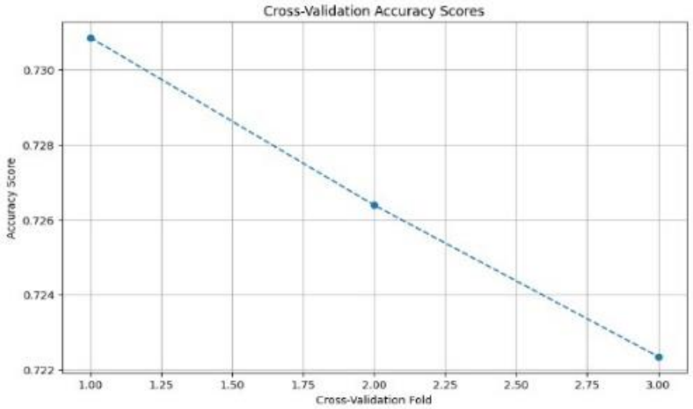

Figure 5: The exponentially increasing graph between the accuracy levels and the cross-validation score graph is being shown as a result of the K-Nearest Neighbors

The above figure 5 depicts the increasing graph between the accuracy and cross validation values.
Long short-term memory was implemented and obtained Test Accuracy is: 0.3356062173843384.

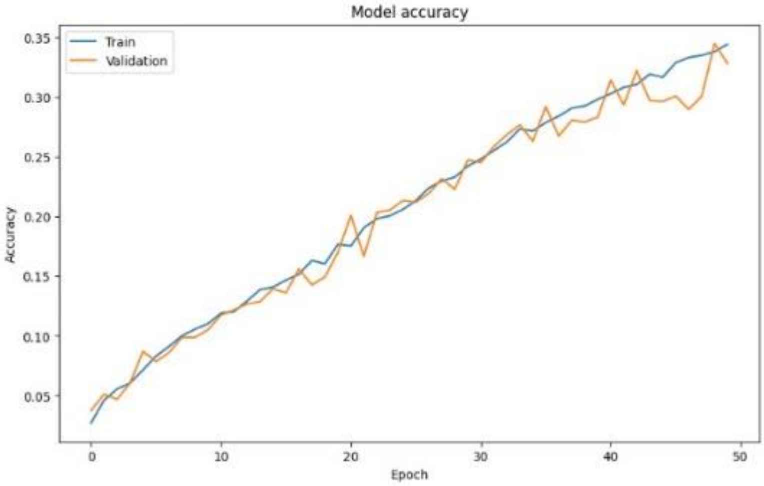

Figure 6: The graph shows the model accuracy obtained by implementing Long short-term memory algorithm.

The above Figure 6 shows the model accuracy rate obtained by implementing Long short-term memory algorithm,

Logistic Regression implemented with the same testing and training variables and the Test accuracy is 0.04005966332836139 and Train accuracy is 0.041822056473095365.

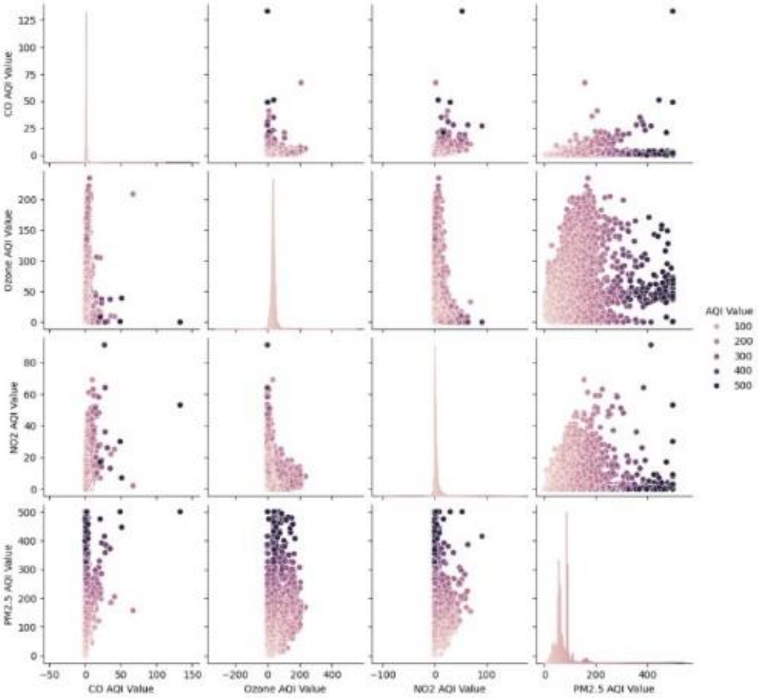

Figure 7: The graph shows the impact various pollutants like Ozone, Particulate Matter 2.5 , Nitrogen Dioxide, Carbon Monoxide on the overall Air Quality Index

The above Figure 7 shows the impact and the effect of major pollutants like Ozone, Particulate Matter 2.5, Nitrogen Dioxide, Carbon Monoxide on the overall Air Quality Index on various cities around the world thus shows the impact of the pollutants on the health of the people in various ways

Table 5
Result and Accuracy scores of the dataset implemented using KNN

Model	Accuracy level
Random Forests.	0.99
K-Nearest Neighbor	0.83
Logistic Regression	0.040
Long short-term memory	0.3356

6. Conclusion

The two types of pollution, such as indoor and outdoor air pollution, are briefly introduced to us in this paper. Also, the effect of air pollution on people affected by Covid-19 Literature review gives us insight on various air pollutants ranging from aerosols like Particulate Matter and major other air pollutants like oxides of Nitrogen and some of the prevalent biological oxides. Each of the pollutants has been briefly explained and its effect on various major cities around the world. For a wider analysis on the Air Quality Index , the data set on Global air quality index is referred and various statistical analysis are done using various tests focusing on the overall air quality index of a city with the amount of Carbon Monoxide , Ozone and Particulate Matter 2.5 present in the air This study used a combination of statistical analyses and machine learning models to examine the effects of ozone and other pollutants on the Air Quality Index (AQI). The main conclusions show that ozone significantly affects AQI, with models such as Random Forest and LSTM best illustrating this effect. While the LSTM caught the temporal patterns of changes in air quality, the Random Forest model highlighted ozone as a critical contributor among other contaminants. KNN offered information on regional pollution trends, while Logistic Regression was a useful standard for classification work. These conclusions were further supported by the statistical testing. Significant variations in AQI levels according to pollutant concentrations were shown by ANOVA findings, and significant relationships between pollutant thresholds and AQI categories were found by cross-tabulation studies. These findings support the machine learning objectives and provide a thorough understanding of the ways in which certain pollutants, especially ozone, impact air quality. There are two ways that this study has ramifications. First of all, the results highlight how crucial it is to continuously monitor and control ozone levels in order to enhance air quality. Second, the efficiency of fusing statistical analysis with machine learning offers a solid framework for evaluating the effects of air pollution that can be extended to additional contaminants and geographical areas. This research provides a great insight into the effect of various air pollutants on the global air pollution trends. Several directions are suggested for further investigation. Increasing the dataset's size to incorporate more areas and contaminants may improve the results' generalizability. Deeper insights might be obtained by incorporating more sophisticated machine learning algorithms and taking other environmental aspects into account. Resolving the limitations that have been highlighted, such as possible biases in the data and the requirement for more detailed temporal data, will enhance the study and make air quality management plans more successful.

References

[1] Tran, V.V., Park, D., Lee, Y.C.: Indoor air pollution, related human diseases, and recent trends in the control and improvement of indoor air quality. Int. J. Environ. Res. Public Health 17(8), 2927 (2020).

[2] Sathe, Y., Gupta, P., Bawase, M., Lamsal, L., Patadia, F., Thipse, S.: Surface and satellite observations of air pollution in India during COVID-19 lockdown: Implication to air quality. Sustain. Cities Soc. 66, 102688 (2021).

[3] Ekmekcioglu, D., Keskin, S.S.: Characterization of indoor air particulate matter in selected elementary schools in Istanbul, Turkey. Indoor Built Environ. 16, 169–176 (2007).

[4] https://www.who.int/news-room/fact-sheets/detail/household-air-pollution-and-health

[5] Tiotiu, A.I., Novakova, P., Nedeva, D., Chong-Neto, H.J., Novakova, S., Steiropoulos, P., Kowal, K.: Impact of air pollution on asthma outcomes. Int. J. Environ. Res. Public Health 17(17), 6212 (2020).

[6] Kumar, K.P., Unal, A., Jha Pillai, V., Murthy, H., Niranjanamurthy, M. (eds.): Data Engineering and Data Science: Concepts and Applications. 2023.

[7] Madhurima, S., Mannooparambil, J.M., Kumar, K.P.: The Road to Reducing Vehicle CO2 Emissions: A Comprehensive Data Analysis. In: International Conference on Computational Sciences and Sustainable Technologies, pp. 299–309. Springer, Cham (2023).

[8] Ashok, K., Reddy, P.A., Kumar, K.P.: ML Algorithms and Their Approach on COVID-19 Data Analysis. In: Kumar, K.P., Unal, A., Jha Pillai, V., Murthy, H., Niranjanamurthy, M. (eds.) Data Engineering and Data Science: Concepts and Applications, pp. 335–349. Springer (2023).

[9] Satpute, B.S., Bharati, R., Rahane, W.P.: Predictive Modeling of Vehicle CO2 Emissions Using Machine Learning Techniques: A Comprehensive Analysis of Automotive Attributes. In: 2023 3rd International Conference on Technological Advancements in Computational Sciences (ICTACS), pp. 511–516. IEEE (2023).

[10] Sadriddinovich, J.T.: Capabilities of SPSS Software in High Volume Data Processing Testing. Am. J. Public Diplomacy Int. Stud. 1(9), 82- (2023).

[11] Sicard, P., Agathokleous, E., Anenberg, S.C., De Marco, A., Paoletti, E. and Calatayud, V., 2023. Trends in urban air pollution over the last two decades: A global perspective. Science of The Total Environment, 858, p.160064.

[12] L. Sushama, K. Sridhar, and M. Roberts, “Deep Learning-based Precision Diagnosis of Lung Diseases on the Internet of Medical Things (IoMT),” Proceedings of the Bulgarian Academy of Sciences, vol. 76, no. 10, Oct. 2023, doi: 10.7546/crabs.2023.10.07.

[13] Yang, J. and Zhao, Y., 2023. Performance and application of air quality models on ozone simulation in China–A review. Atmospheric Environment, 293, p.119446.