A Study the Impact of household air pollution on human health using Machine Learning Techniques.

Prakash S. Chougule [1], Tejaswi S.Kurane[2] , Shubham D.Shedge[3] , Ramesh D.Shinde[4] , Rahual. H. Waliv[5] ,Mrs.Varsha. C.Shinde[6] ,Rajnikant G.Jawalkote[7]

[1] Associate Professor, Rajarshi Chhatrapati Shahu College, Kolhapur(MS), India

[2,3]Assistant Professor, Rajarshi Chhatrapati Shahu College, Kolhapur (MS), India

[4]Assistant Professor, Jaysingpur College, Jaysingpur (MS), India

[5]Associate Professor, Kisan Veer Mahavidyalaya, Wai Dist.Satara (MS), India

[6] Associate Professor, Vivekanand College, Kolhapur(MS), India

[7] Research Student, Rajarshi Chhatrapati Shahu College, Kolhapur (MS), India

Abstract: Due to the rapid development of technology, urbanization and increased population, air pollution has become a hot topic, in particular because of the effects on health. However, much of the focus has been on outdoor air pollution as well as indoor air pollution , Some of the most important sources of indoor air pollution are Volatile Organic Compounds (VOCs) and Particulate Matter (PM). There are a variety of VOCs emitted from modern household products (e.g., paints, lacquers, cleaning liquids, furnishings, copiers, printers, glues, adhesives or permanent markers). Air pollution is a major environmental health threat. Exposure to fine particles in both the ambient environment and in the household causes about seven million premature deaths each year. The main indoor air pollutants are Volatile Organic Compounds (VOCs) and Particulate Matter (PM). PM sources included smoking, cooking, heating, candles, and insecticides, whereas sources of coarse particles were pets, housework and human movements. VOC sources included household products, cleaning agents, glue, personal care products, building materials and vehicle emissions. This public health crisis is receiving more attention, but one critical aspect is often overlooked: how air pollution affects children in uniquely damaging ways. Recent data released by the World Health Organization (WHO) show that air pollution has a vast and terrible impact on child health and survival. In this study we take secondary data of India was taken from git hub. Analyses the data by using machine learning models like Random search CV, Bagging Classifier, Linear regression, logistic regression, XG Boost . for comparing the all the factors of AQI values useing linear regression ,logistic regression and XG Boost models .it is observed that XG Boost has high accuracy than linear regression ,logistic regression and logistic shows better performance as compared to linear regression and it is shows that for PM2.5 values randomized search CV and for AQI values of all factors XG Boost and logistic regression are fitted good.

Keywords: Bagging Classifier, Randomized Search CV, Linear regression, logistic regression, XG Boost ,Public Health, Air Quality Index, and Household Air Pollution.

Introduction: Pollution is becoming an alarming threat to our planet day after day. Food pollution has been the focus of national and international public health organisations, particularly pesticide residues and bioaccumulating substances. They have also focused on reducing outdoor air pollution caused by cities, factories, and automobile exhaust emissions. Meanwhile, whereas people in high-income countries (HICs) spend much of their lives indoors, the pollution of the indoor environment still needs to be addressed [1–3]. Indeed, domestic air and indoor pollution can be traced back to prehistory, when humans first moved to temperate climates, started building shelters, and used fire for cooking, heating, and lighting. Indoor pollution is a global health issue. Today, all over the world, about 2.4 billion people still make food with solid stuff (like wood, farm leftovers, coal, and animal poop). Many of these individuals are impoverished and reside in low- and middle-income nations, with a significant gap between urban and rural settings. In 2020, just 14% of people living in urban areas depended on dirty fuels and outdated technologies, in sharp contrast to the 52% rate of the global rural population.[4]. Despite transitioning from biomass fuels to petroleum products and electricity accompanying modernisation in developed countries, pollution remains a persistent threat to public health [5]. Although inhalation is the primary way indoor pollutants are exposed, it is important to consider cutaneous and oral exposure, especially for children who frequently interact with their hands and frequently participate in activities that involve contact with floors [6][7]. According to Wilson's research, kids touch their mouths, eyes, and noses more often than adults do. In particular, hand-to-mouth contact may be

a matter of concern when considering exposures to chemicals, such as lead or pesticides [6]. Sources of HAP are numerous and can be biological, chemical or physical. Solid fuels (e.g., coal, biomass, and animal dung) are the leading sources of HAP; these fuels are used by 41% of households and approximately half of the world's population as their principal household fuel for cooking, heating, and lighting [8][9]. The use of these solid fuels has become a significant public health problem and is attracting great attention. Indeed, household use of solid fuels was calculated to be among the top five major risk factors for global disease in 2010 (4.3.% of global disability-adjusted life-years (DALYs), 95% confidence interval (CI) = 3.4–5.3%), after tobacco smoking [10]. Suppose we focus on HAP caused by the use of biomass fuels-. In that case, we can identify many health threats strongly responsible for respiratory tract infections, the potency of inflammatory lung conditions, cardiac problems, stroke, eye disease, tuberculosis (TB), and cancer. Moreover, when collecting firewood to use as fuel, women and girls are threatened by numerous indirect health effects, not only because they carry large bundles of wood on their heads and necks but also because they must travel far from home, which car rise risks such as assault, insects (which are disease vectors), snake bites, absence from school or other learning opportunities, and musculoskeletal injuries [11]. Second, personally carried samplers were used to sample air pollutants, most of which are particulate matters (PMs) and the pollutants adsorbed in PMs, which could not present information about the real-time exposure [12, 13, 14]

Main risk factors concerning indoor air and environment: Indoor air and environment risk factors Certain individual Factors that pose risks associated with the indoor environment. are quite well-recognized, for example, airborne infections (air conditioning), pathogens and allergens, radon, and passive tobacco smoking; however, the picture concerning indoor risk factors is much more complex and requires a thorough evaluation. The expert group identified the significant indoor air quality risk factors that need further research: Ventilation, climate factors, chemistries, and socio-economic status.This article surveyed the sources and types of air pollutants inside households and their harmful effects on human health. It[this culture of HAP] seeks to provide a strong foothold for the pursuit of knowledge regarding HAP as a worldwide health problem. We finally provide an empirical viability analysis of management practices for source elimination of household air contaminants through various approaches.

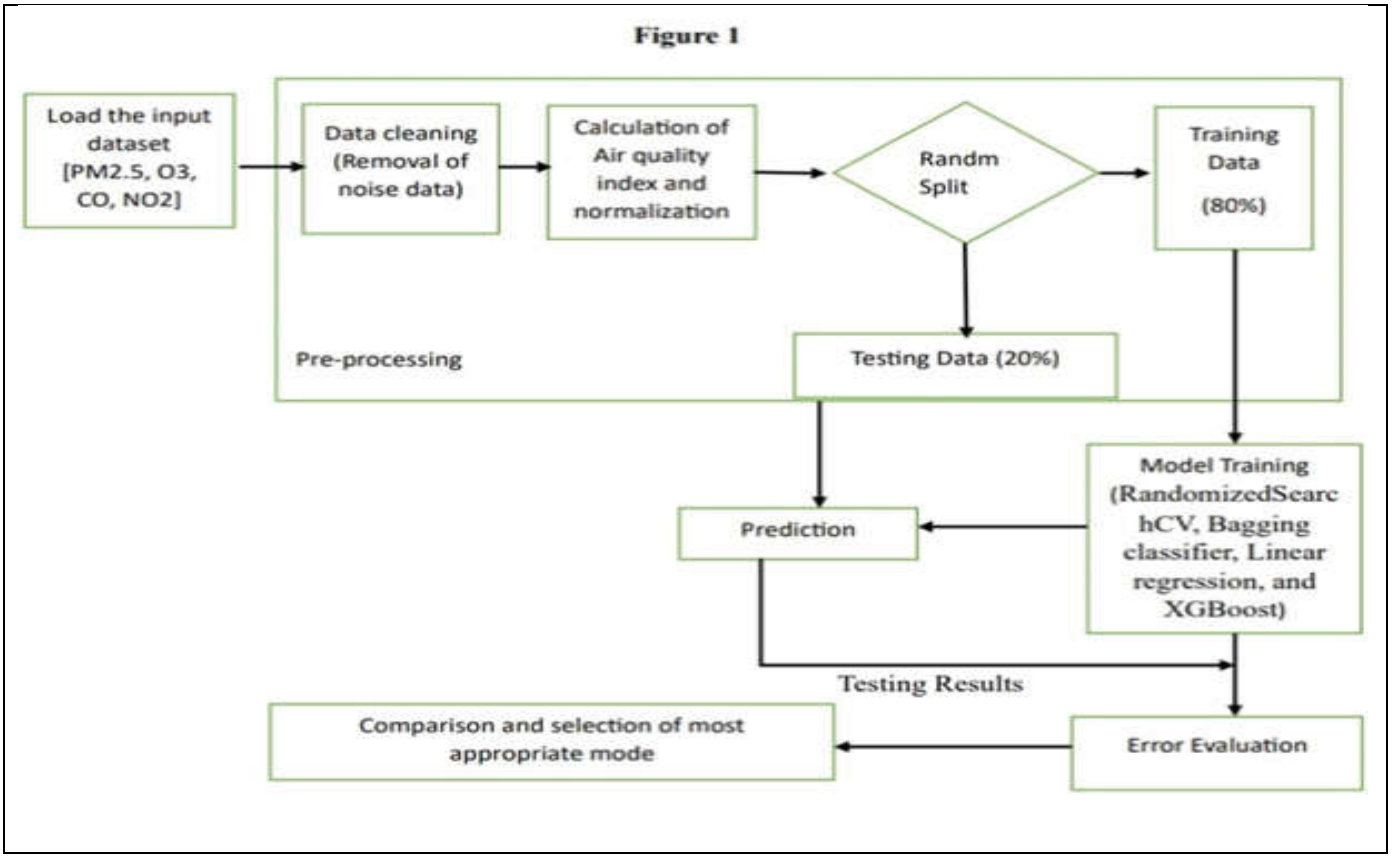
Literature review: In a research study, the mentioned regression model was implemented to estimate PM10 concentration (target) in Conjure, Thailand, with the help of independent input data, meteorological and pollutants. The meteorological inputs included air pressure, precipitation, temperature, relative humidity, and wind speed. The pollutants included carbon monoxide (CO), nitrogen monoxide (NO), nitrogen dioxide (NO2), sulphur dioxide (SO2), Black Carbon (BC), methane (CH4), Non-Methane Hydro Carbon (NMHC) and ozone (O3) from 2006 to 2008 (Saithanu & Mekparyup, 2014)[15]. Another study by Rybarczyk and Zalakeviciute estimated PM2.5 concentration with regression models based on time. Primarily, the regression models were built by segregating the day into three-time segments such as 6 a.m.–10 a.m., 10 a.m. to 2 p.m. and 2 p.m.–7 p.m. for the capital city of Quito, Ecuador. In this study, the models were initially built for each time period based on ease of data availability. Traffic was considered for the three periods and segregated into high, medium and low. The model showed an R2 score of 0.27 with these settings, adding meteorological data, including solar radiation, air temperature, air pressure, precipitation, relative humidity, and wind speed as features, improving the R2 score to 0.38. Finally, the trace gas concentrations SO2, NO2, O3 and CO were also considered, which improved the R2 score to 0.8. The limitation of this study was the extra cost associated with measuring the trace gas concentrations (Rybarczyk & Zalakeviciute, 2017) [16]. Due to their broad applicability and popularity, the ensemble methods have been used in air pollution estimation. Most studies using this approach used meteorological parameters such as air temperature, air pressure, relative humidity, and wind speed as input parameters. These parameters vary with location and are crucial in the rapid changes in pollutant concentrations. PM2.5 forecasting was performed on a single monitoring station, including the meteorological parameters mentioned in Delhi, India. Overall, eleven models were utilised, and the R2 scores were compared. However, in this study, the outputs from two different models were combined to see the improved performance [17]. In a study in Munich, Germany, the XGBoost model was built using meteorological parameters, precursors and simulations of ozone concentration obtained from the CAMS2 dataset to estimate ozone concentration. This study aimed to investigate the significance of precursor information in modelling surface ozone using ML. The meteorological parameters include factors like air temperature and relative humidity, boundary layer height, wind speed and wind direction, as well as in-situ ozone precursors (NO, NO 2 cursors (column NO 2 and CO), and satellite ozone pre and HCHO) along with CTM simulations (CAMS model surface O 3) were used as input parameters. The day of the week and season were also considered [18]. Satellite-based estimates of daily NO 2 exposure in China were tested using a

hybrid random forest and spatiotemporal rigging model [19]. A random forest model for ozone estimation was built at the Research Academy for Environmental Sciences in Beijing, China [20].

Methodology: This study assesses four distinct models utilising various machine learning techniques, such as Randomized Search CV, Bagging classifier, linear regression, and XG Boost (Extreme Gradient Boosting). We have compared these models to determine the most suitable one for effectively forecasting the air quality index (AQI). These methodologies will cover the collection and loading of air quality datasets, data pre-processing, building models, training the models, making predictions, model performance evaluation, and the most appropriate model selection. The air quality data was initially collected for analysis. The study consisted of cleaning noise data, computing the air quality index (AQI), and dividing it into training and testing sets. The processed data contained information on PM2.5, CO, NO2, O3 and their AQI values over 24 hours. The models were successfully trained on the pre-processed data for the testing step; we assessed the models using test set pollutant data to check how accurately they could predict AQI. We then compared observed and predicted data to produce performance metrics whereby we could analyse each predictive model and determine the most appropriate one for making accurate predictions of AQI. In this study, we looked at data on air quality collected from ten Indian cities. The information gathered from the website GitHub.Com included several important variables that were thought to be required for precisely forecasting the air quality index (AQI), including PM2.5 (particle diameter ≤ 2.5 µm), NO2, O3, and CO.

A] Data processing: Any data analysis project must include data pre-processing, in which we attempt to identify and remove noise from the datasets. Data cleaning, AQI calculation, and dataset division into training and testing sets are the three main sub-steps in this step. We start by filling in any missing values in the raw data. We used a simple imputation technique with a central tendency measure to fill these gaps. The median is typically a more reliable option than the mean when there are outliers in the data. Thus, we substituted median values for the missing data for several pollutant parameters, including PM2.5, O3, CO, and NO2. Additionally, we found and eliminated outliers and duplicate entries. Next, we used various air pollutant parameters from a few chosen stations throughout India to compute AQI values. The AQI is essential for assessing air quality in a given location. The computed PM2.5 and AQI values were added to the dataset as target variables. We normalised the data to improve the learning efficacy of the model and account for the significance of various variables. This study used min-max normalisation—which linearly transforms the data. Separate normalisations were performed for O3, NO2, CO, and PM2.5. This method preserves the relationships between the various variables while scaling the values to a range between 0 and 1 to maintain the data patterns. A methodical flow diagram in Figure 1 shows the method for predicting air quality.

Figure 1



B] Train and test split:

The available hourly data was separated into training and test data following feature identification. They chose the training data. Twenty per cent of the data was used for testing, and the remaining eighty per cent was used for training. The train and test split percentages were based on the literature review, in which many authors suggested the division of this range. A 5-fold cross-validation technique was employed, i.e., dividing the train set into five equal parts and, each time, training four parts and evaluating the remaining part as the validation set. Thus, in this way, the generalising ability of the model increases while training and, hence, when evaluated on the test, may yield plausible results. This form of evaluation, on one part, after training on nine parts, is used to find out the best set of hyper parameters for each of the models.

C] Metrics: The predicted Air quality index was measured using five metrics, namely, precision (Prec), recall (Rec), f1score (F1), and accuracy (Acc.) [21]. Let the letter be: Tp = True positive or occurrences where the model predicted the positive air quality index truly, Tn = True negative or occurrences where the model predicted the negative class truly, Fp = False positive or occurrences where the model predicted the positive class falsely, Fn = False negative or occurrences where the model predicted the negative class falsely, Precision, recall, accuracy, and f1score shown in equations given below,

$$Precision = \frac{Tp}{Tp + FP}$$

$$Recall = \frac{Tp}{Tp + Fn}$$

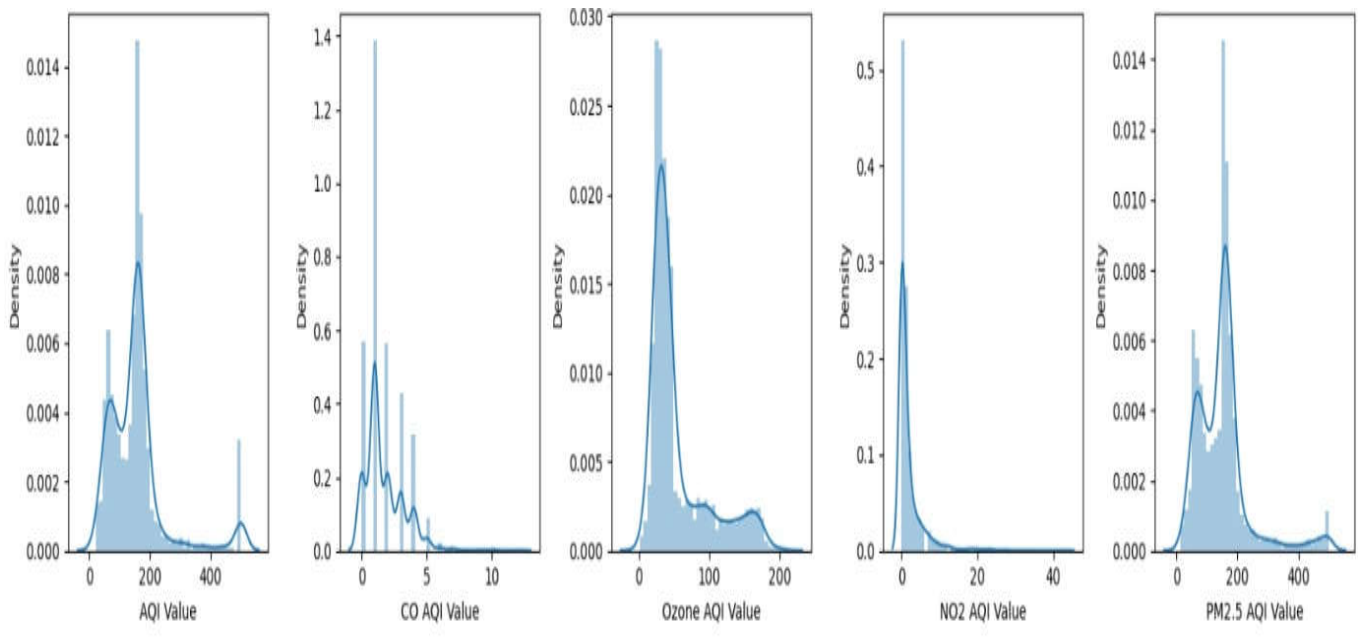
$$Accuracy = \frac{Tp + Tn}{Tp + Tn + Fp + Fn}$$

$$F1score = 2.\frac{Precision.Recall}{Precision + Recall}$$

The area under the curve (Auc) score helps distinguish a classifier's capacity to compare classes and is utilised to review the region operating curve (roc). The Roc curve visualises the relationship between true positive rate and false positive rate across various thresholds.

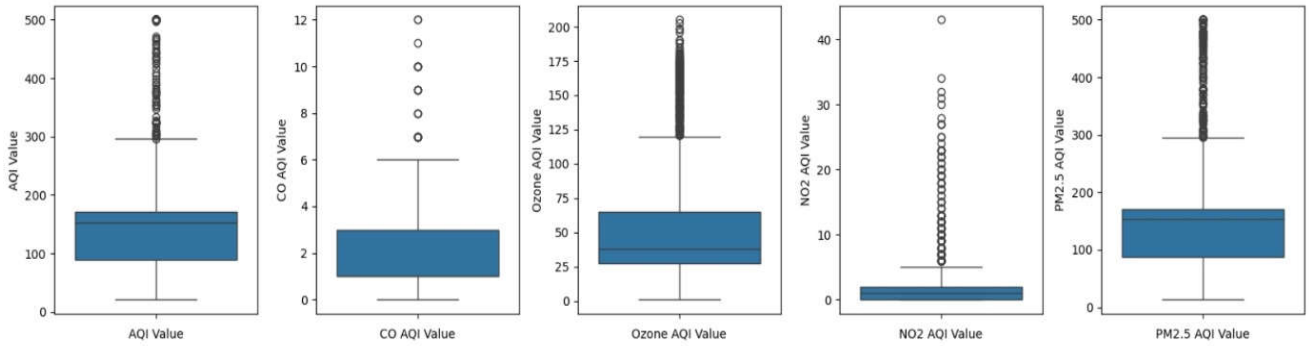
Statistical analysis:

Figure 2



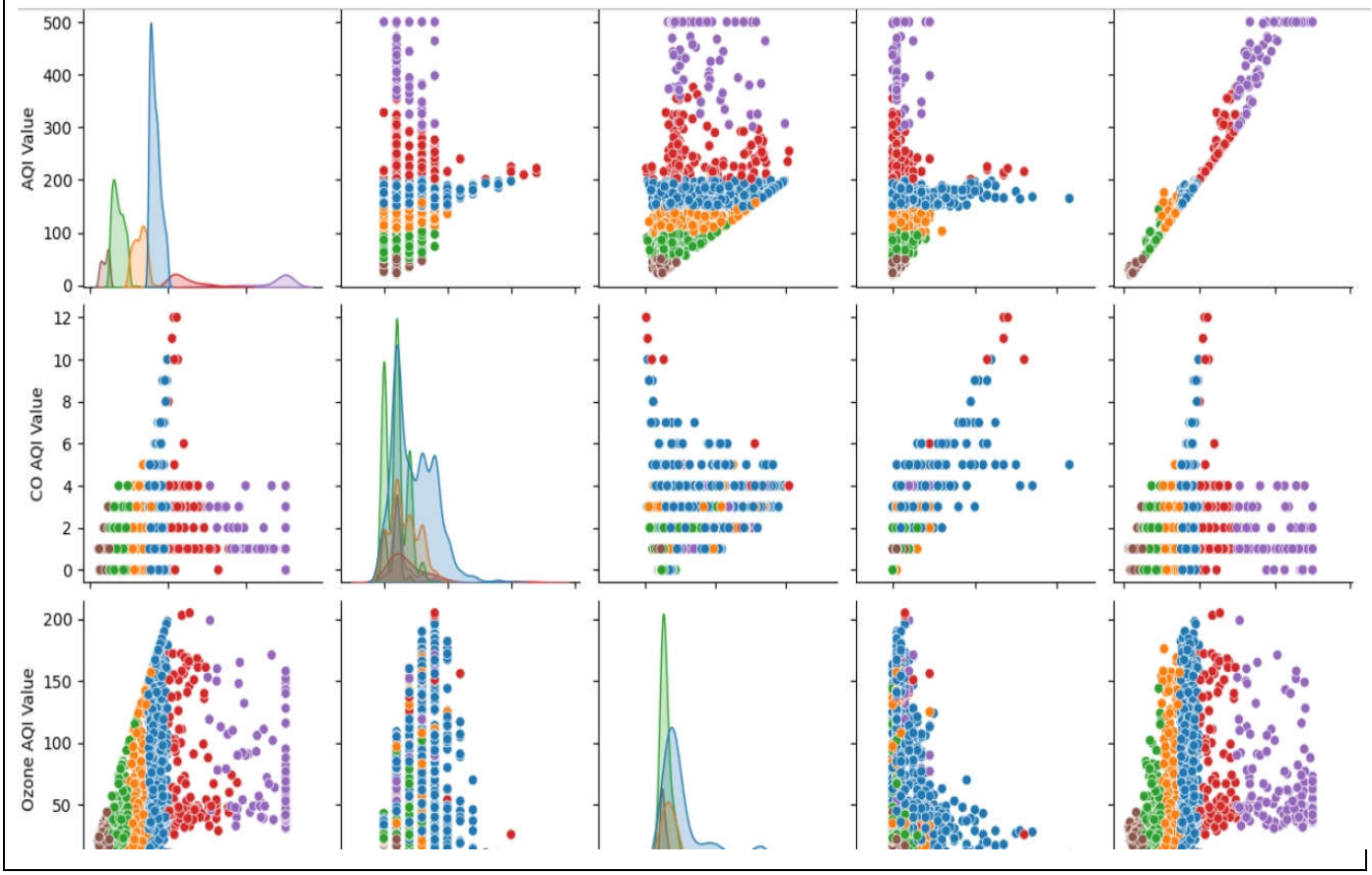
It displays the distribution of AQI values (e.g. G. multimodal, skewed, or standard). It detects extreme values or possible outliers. It helps compare different pollutants' AQI distributions. It provides insights into pollution trends, peaks, and variations.

Figure.3



The circles above the whiskers represent extreme values that are significantly different from the majority of data points. For example, the AQI values of ozone, NO2, and PM2.5 have many outliers.

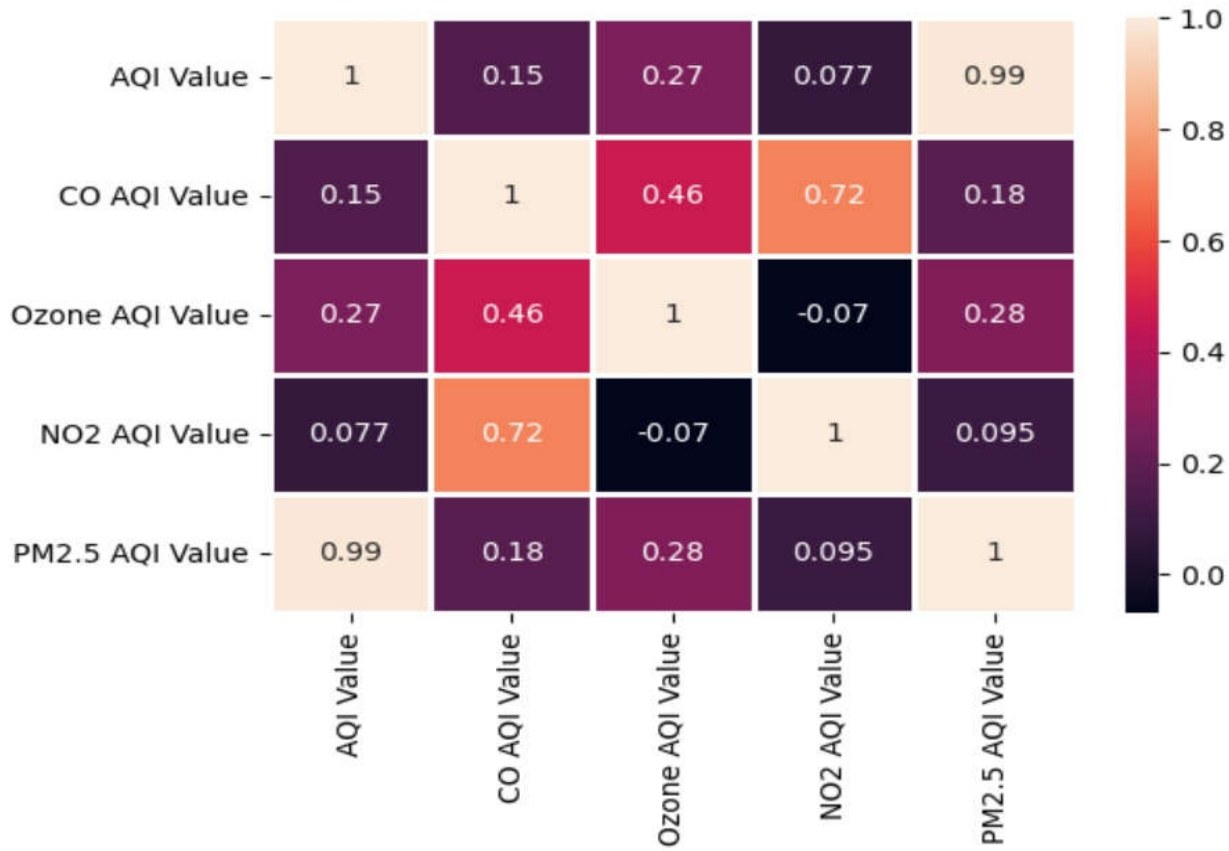
Figure.4



Each scatter plot shows how two pollutants are related. If points align in a diagonal pattern, it suggests a strong correlation (e.g., AQI vs. PM2.5). There is no strong correlation if points are scattered randomly (e.g., CO vs. Ozone).Using this spatial information, stations at different sites can assist in estimating the pollutants of interest, PM 2.5 and NO 2, at the two locations Marienplatz and Am Neckartor. The pollutant concentration is affected by certain factors not considered for ML models to make it less complex, such as the location characteristics, topography, varying emission sources, and occasional activities, e.g. construction, festivals, etc. One of the reasons for including pollutant concentration from other stations as an input was to consider such factors indirectly by using the pollutant concentration of other stations. A Spearman rank correlation matrix shows the relationship between pollutants measured at different monitoring stations [22]. In Fig. 2.

This plot Measures the relationship between two variables in order of their ranks. Thus, it essentially measures the monotonic relationship between those two variables.

Figure.5



All statistical analyses were performed using Python. Data were summarised descriptively with various statistics. Skewness statistics and normal probability plots were used to assess continuous variables' normality. The new users of solid fuels or the chronic clean fuel users were considered the reference group. Duration of solid fuel use in cooking was our main exposition of interest, which was categorised at four levels. We adjusted a significant exposition decline to use solid fuels to cook during the last 12 months with linear regression. In this secondary data analysis research, 80% of the data of the 2,488 individual participants were utilised for training, while the other 20% was held for inspection.

Table 2

Model	Precision	Recall	F1-score	Accuracy
Randomized Search CV	0.88	0.92	0.90	0.85
Bagging	0.87	0.89	0.88	0.83
Linear regression	0.92	0.83	0.87	0.87
XG Boost	1	0.98	0.99	0.99
Logistic Regression	0.99	0.98	0.99	0.99

Conclusion:

For this study we take secondary data of India was taken from git hub. We are clean the data by remove the absent data then the coding and classification is done. Then we make the normal data for this purpose we use AQI value and normality test. it is observed that some factors are correlated and pattern of data is right skewed .then we identified and removed out the extreme points(outliers) in AQI values. Ozone, No2 and PM2.5 factors includes more outliers, from Hit map we observed that AQI value and PM2.5 AQI value factors are high correlated. The we check performance of Random search CV, Bagging Classifier, Linear regression, logistic regression, XG Boost models for above data. For the target variable PM2.5 AQI value , in Random search CV model decision tree classifier and cross validation of five folds are used. Randomized search CV shows excellent performance than bagging classifier and random search CV has more accuracy than Bagging Classifier. In the absence of PM2.5 AQI values we used three models like linear regression, logistic regression and XG Boost comparing the all the factors of AQI values used three models .it is observed that XG Boost has high accuracy than linear regression ,logistic regression and logistic shows better performance as compared to linear regression. From our study it is shows that for PM2.5 values randomized search CV and for AQI values of all factors XG Boost and logistic regression are fitted good.

References:

1. Air pollution and child health: Prescribing clean air. (n.d.). Retrieved March 01, 2025, from https://www.who.int/publications/i/item/WHO-CED-PHE-18-01

2. Household air pollution. (n.d.). Retrieved March 02, 2025, from https://www.who.int/news-room/fact-sheets/detail/household-air-pollution-and-health

3. IEA IRENA, World Bank UNSD, Tracking WHO. SDG 7: The Energy Progress Report. World Bank, Washington DC. © World Bank. License: Creative Commons Attribution Non -commercial 3.0 IGO (CC BY-NC 3.0 IGO).

4. Wilson, A. M., Verhougstraete, M. P., Beamer, P. I., King, M.-F., Reynolds, K. A., & Gerba, C. P. (2021). Frequency of hand-to-head, -mouth, -eyes, and -nose contacts for adults and children during eating and non-eating macro-activities. Journal of Exposure Science & Environmental Epidemiology, 31(1), 34–44.

5. Maung, T. Z., Bishop, J. E., Holt, E., Turner, A. M., & Pfrang, C. (2022). Indoor air pollution and the health of vulnerable groups: A systematic review focused on particulate matter (PM), volatile organic compounds (VOCs) and their effects on children and people with pre-existing lung disease. International Journal of Environmental Research and Public Health, 19(14), 8752.

6.Bruce, N., Perez-Padilla, R., & Albalak, R. (2000). Indoor air pollution in developing countries: A major environmental and public health challenge. Bulletin of the World Health Organization, 78(9), 1078–1092.

7. Bonjour, S., Adair-Rohani, H., Wolf, J., Bruce, N. G., Mehta, S., Prüss-Ustün, A., Lahiff, M., Rehfuess, E. A., Mishra, V., & Smith, K. R. (2013). Solid fuel use for household cooking: Country and regional estimates for 1980–2010. Environmental Health Perspectives, 121(7), 784–790.

8. Lim, S. S., Vos, T., Flaxman, A. D., Danaei, G., Shibuya, K., Adair-Rohani, H., AlMazroa, M. A., Amann, M., Anderson, H. R., Andrews, K. G., Aryee, M., Atkinson, C., Bacchus, L. J., Bahalim, A. N., Balakrishnan, K., Balmes, J., Barker-Collo, S., Baxter, A., Bell, M. L., … Ezzati, M. (2012). A comparative risk assessment of burden of disease and injury attributable to 67 risk factors and risk factor clusters in 21 regions, 1990–2010: A systematic analysis for the Global Burden of Disease Study 2010. The Lancet, 380(9859), 2224–2260.

9. Oluwole, O., Otaniyi, O. O., Ana, G. A., & Olopade, C. O. (2012). Indoor air pollution from biomass fuels: A major health hazard in developing countries. Journal of Public Health, 20(6), 565–575.

10. Huang, Y., Du, W., Chen, Y., Shen, G., Su, S., Lin, N., Shen, H., Zhu, D., Yuan, C., Duan, Y., Liu, J., Li, B., & Tao, S. (2017). Household air pollution and personal inhalation exposure to particles (TSP/PM2.5/PM1.0/PM0.25) in rural Shanxi, North China. Environmental Pollution, 231, 635–643.

11. Hu, W., Downward, G. S., Reiss, B., Xu, J., Bassig, B. A., Hosgood, H. D., Zhang, L., Seow, W. J., Wu, G., Chapman, R. S., Tian, L., Wei, F., Vermeulen, R., & Lan, Q. (2014). Personal and indoor PM2. 5 exposure from burning solid fuels in vented and unvented stoves in a rural region of China with a high incidence of lung cancer. Environmental Science & Technology, 48(15), 8456–8464.

12. Wu, C., Li, Y.-R., Kuo, I.-C., Hsu, S.-C., Lin, L.-Y., & Su, T.-C. (2012). Investigating the association of cardiovascular effects with personal exposure to particle components and sources. Science of The Total Environment, 431, 176–182.

13. Saithanu, K., Mekparyup, J., 2014. Using multiple linear regression to predict pm 10 concentration in Chonburi, Thailand. Global Journal of Pure and Applied Mathematics, Bd. (10), 835–839, 122014.

14. Rybarczyk, Y., & Zalakeviciute, R. (2018). Regression models to predict air pollution from affordable data collections. In H. Farhadi (Ed.), Machine Learning—Advanced Techniques and Emerging Applications. InTech. https://doi.org/10.577/intechopen.71848.

15. Analitis, A., Barratt, B., Green, D., Beddows, A., Samoli, E., Schwartz, J., & Katsouyanni, K. (2020). Prediction of PM2.5 concentrations at the locations of monitoring sites measuring PM10 and NOx, using generalized additive models and machine learning methods: A case study in London. Atmospheric Environment, 240, 117757.

16. Balamurugan, V., Balamurugan, V., & Chen, J. (2022). Importance of ozone precursors information in modelling urban surface ozone variability using machine learning algorithm. Scientific Reports, 12(1), 5646.

17. Zhan, Y., Luo, Y., Deng, X., Zhang, K., Zhang, M., Grieneisen, M. L., & Di, B. (2018). Satellite-based estimates of daily no2 exposure in china using hybrid random forest and spatiotemporal kriging model. Environmental Science & Technology, 52(7), 4180–4189.

18. Zhan, J., Liu, Y., Ma, W., Zhang, X., Wang, X., Bi, F., Zhang, Y., Wu, Z., & Li, H. (2022). Ozone formation sensitivity study using machine learning coupled with the reactivity of volatile organic compound species. Atmospheric Measurement Techniques, 15(5), 1511–1520.

19.Powers, David & Ailab,. (2011). Evaluation: From precision, recall and F-measure to ROC, informedness, markedness, and correlation. J. Mach. Learn. Technol. 2. 2229-3981.

20. Akoglu, H. (2018). User's guide to correlation coefficients. Turkish Journal of Emergency Medicine, 18(3), 91–93.